Semantic Change Characterization with LLMs using Rhetorics

Anonymous ACL submission

Abstract

Languages continually evolve in response to societal events, resulting in new terms and shifts in meanings. These changes have significant implications for computer applications, including automatic translation and chatbots, making it essential to characterize them accurately. The recent development of Large Language Models (LLMs) has notably advanced natural language understanding, particularly in sense inference and reasoning. In this paper, we investigate the potential of LLMs in characterizing three types of semantic change: dimension, relation, and orientation. We achieve this by combining LLMs' Chain-of-Thought with rhetorical devices and conducting an experimental assessment of our approach using newly created datasets. Our results highlight the effectiveness of LLMs in capturing and analyzing semantic changes, providing valuable insights to improve computational linguistic applications.

1 Introduction

Language, a tool humans acquired throughout evolution, remains a subject of fascination and inquiry across diverse disciplines, including neuroscience, psychology, philosophy (Pinker, 2003), and computational linguistics. Despite this interdisciplinary interest, our understanding of language is often superficial, with much to uncover regarding its intricacies (Allan, 2013; Pinker and Bloom, 1990). Among the many elements that shape language, a central aspect in understanding dynamics in language development is how the semantics of words change (Allan, 2013; Pinker and Bloom, 1990). This evolution is particularly intriguing in computational linguistics, as it impacts applications such as automatic translation and chatbots (Camboim de Sá et al., 2024). While humans can rapidly adapt to changes using a lot of contextual information and cognitive processes to grasp the senses of a sentence or a word, it is complex to provide enough cultural knowledge and nuances to machines. Consequently, machines lack the tools to adapt to these variations and to perform effective communication (Tahmasebi et al., 2018). Therefore, in many modern Natural Language Processing (NLP) systems, we observe the impacts of semantic change on end users, especially when the task requires deep contextual dependency (Camboim de Sá et al., 2024).

In the context where historical or domain-specific knowledge of meaning is crucial, the Lexical Semantic Change (LSC) field emerged to gain deep understanding of and detect these changes (Tahmasebi et al., 2018). While a significant body of work explores which words changed in different moments or domains, there is still a need for further comprehension regarding the types and implications of semantic changes in these systems (Hengchen et al., 2021). For instance, in sentiment analysis, being aware that the term 'sick' has acquired a positive connotation could significantly alter the interpretation of a sentence.

Theories to comprehend semantic change exist. One prominent typology, proposed by (Traugott, 2017), categorizes change into broadening/narrowing (a word gains or loses senses), amelioration/pejoration (a word is perceived more positively/negatively), metaphorization and metonymization (the word is used as a metaphor or metonymy respectively). We illustrate these types of change in Table 1.

A child in **dirty** overalls. He used a **dirty** trick to win the competition.	pejoration
No other style of **hat** was acceptable with an evening dress. He took off his politician's **hat** and talked frankly.	metaphorization
The diamond is currently set in the **crown** of the Queen. The colonies revolted against the **crown**.	metonymization

Table 1: Examples illustrating the characterization of types of change.

Previous works have only partially covered the typology of semantic change, typically focusing on a few types (Camboim de Sá et al., 2024). However, recent advancements in Language Models (LLMs) have showcased capabilities in executing complex linguistic tasks such as inference, association, understanding, and common sense reasoning (OpenAI, 2023) via the Chain-of-Thought (Chain-of-Thougth (CoT)) technique (Wei et al., 2022). These reasoning abilities in LLMs mimic human-like processes for establishing connections and relations in natural language (Dasgupta et al., 2022).

Rhetorical devices, known for their role in building persuasive arguments and reasoning, utilize cognitive processes to improve argumentation and communication efficiency (Lakoff and Johnson, 2008). These devices facilitate concise communication and also exhibit characteristics related to semantic evolution over time. Hence, they have been extensively employed by human evaluators to compare senses (Kearns, 2006; Steen et al., 2007). In this process, evaluators use historical and cultural knowledge to explain variations in semantic change through rhetorical argumentation.

Moreover, recent studies have found that LLMs encode extensive cultural knowledge, including relationships, associations, and events (Petroni et al., 2019), making them suitable for automating the characterization of semantic change. Building upon this insight, we propose leveraging rhetoric in natural language to characterize semantic change within the proposed typology by exploring LLMs "thought" processes to mimic human cognitive reasoning. Following the outlined typology, our methodology aims to instruct LLM to utilize rhetorical devices and characterize change within a comparative framework. Our contributions are:

- A new approach to semantic change characterization exploring "reasoning" and rhetoric capabilities of LLMs.
- The proposal of 3 new public datasets for evaluation of semantic change characterization: dimension, orientation, and relation.

The paper is structured as follows: Section 2 presents related work of the field semantic change characterization using LLMs. Section 3 details our methodology for prompting models for semantic change identification and characterization. Section 4 introduces the experimental settings, including datasets and results. Section 5 discusses insight from the method. Finally, Section 7 contains concluding remarks and outlines future work.

2 Related Work

Most of the papers in semantic change address the problem of identification, i.e., detecting if the meaning of a word changed without inferring what type of change occurred. In the context of Language Models (LMs) some authors explore these capabilities to track semantic change identification as a sequence-to-sequence problem (Lyu et al., 2022; Giulianelli et al., 2023), by first prompting the model to disambiguate the word in context and then generating a contextualized word representation.

In challenges for semantic change identification for Russian and Spanish (Pivovarova and Kutuzov, 2021; Zamora-Reina et al., 2022), the best performing methodologies were large cross-language models fine-tuned in Word Sense Disambiguation (WSD) for English data to then fit a linear regression over the contextualized embeddings to identify semantic change. Later, LLMs were employed for this task (Wang and Choi, 2023; Periti et al., 2024), but using only few-shot prompt or fine-tuning to the task.

Previous works on semantic change characterization relies on extracting word representations for each corpus, and later compare them to capture possible differences in usage (Camboim de Sá et al., 2024). In the context of broadening/narrowing Bochkarev et al. (2022) utilize a neural network to determine if a word is employed as a named entity. This approach creates a temporal perspective of a word's usage and allows them to compare occurrences of a word to see whether it has gained new usage in the corpus. For metaphorization Maudslay and Teufel (2022) fine-tuned with supervision a BERT model to classify contextualized words into metaphor and then analyzed different corpus. Finally, in the amelioration/pejoration context Fonteyn and Manjavacas (2021) measures polarity in the term 'to death' by calculating the distance between the word vectors 'good' and 'bad'.

Compared to previous works on semantic change, this is the first study to use CoT for this task, with our approach being deeply motivated by linguistic literature. In terms of semantic change characterization, this is the first work that generalizes across all types of change (Camboim de Sá et al., 2024), has no dependency on training data, and can be used for every type of relation e.g., metaphor and metonymy.

3 Methodology

3.1 Background

In this paper, we propose a method for automating the characterization of semantic change across different corpora. To this end, we rely on the following set of predominant typologies defined in the literature (Traugott, 2017; Juvonen and Koptjevskaja-Tamm, 2016):

- **Broadening:** gaining a new meaning related or not to the previous meaning, such that a word represents more concepts, e.g., 'cloud,' a computing infrastructure.
- **Narrowing:** restriction of meaning occurs when a symbol represents fewer concepts than previously, e.g., 'gay' which historically meant festive or happy, is now predominantly used to refer to homosexuality.
- **Amelioration:** a word gains a more positive sense to the previous sense, *nice*, 'foolish, innocent' changed to 'pleasant.'
- **Pejoration:** the word is used with a worse connotation to the previous usage, *stincan*, 'smell (sweet or bad)' changed to *stink*.
- **Metonymization:** association between terms, e.g., *board* 'table', changed to "people sitting around a table, governing body."
- **Metaphorization:** conceptualizing one thing in terms of another, e.g., 'head of the company' the word 'head' conceptualizes "command or control."

We used the same classification of typologies as presented by Camboim de Sá et al. (2024), where the typologies can be regrouped into three poles, namely Dimension, Relation, and Orientation (see Figure 1).

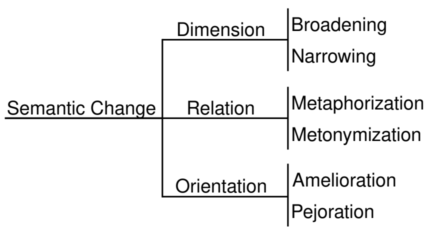

Figure 1: Taxonomy for the poles of Lexical Semantic Change (Traugott, 2017; Juvonen and Koptjevskaja-Tamm, 2016).

In the **dimension** pole, we compute the "number of senses" a word can have. This pole is self-complementary as increasing represents broadening, and decreasing represents a narrowing of senses. After identifying the number of senses, we can compare the differences between corpora.

Metaphorical and metonymical changes are classified under the **relation** category, as these changes enhance the connection between one sense of a word and its other senses. In this framework, a word's meaning relies on the link established through either conceptual (abstract relation) or material (physical association) similarity between concepts. We identify which senses are used figuratively in relation to other senses of the same word.

The **orientation** pole regroups the process of amelioration or pejoration of a meaning. In this pole, words are analyzed according to the contextual sentiment captured from each corpus, and then we analyze how the sentiment changes over corpora. In this study, we explore only positive, negative, and neutral sentiment values for words.

3.2 Rhetorical Arguments as a Pragmatic Tool

LLMs have exhibited significant progress in natural language comprehension. This includes reasoning by analogy (Webb et al., 2022), understanding metaphors (Liu et al., 2022), argumentation (Chen et al., 2023), and acquiring cultural knowledge (Petroni et al., 2019). Additionally, instructing an LLM to generate a rationale, which is a natural language explanation for its reasoning process, before providing an answer has been shown to improve performance on many NLP tasks that require logical reasoning (Wei et al., 2022; Kavumba et al., 2023). This rationale generation step is believed to inject more information retrieved from the LLM's internal knowledge store into the prompt. This enriched prompt allows the LLM to consider a broader range of knowledge during the final decision-making process (Dasgupta et al., 2022).

In this paper, we rely on its stored cultural knowledge to improve the context for the task and work as an annotator in the framework proposed in the previous subsection. The problem of characterizing LSC, identifying different senses, figurative usage, and feelings, relies on building cognitive relations between other senses that depend on human perception and culture (Lakoff and Johnson, 2008). As language is the best tool to explain language (Pinker, 2003), in our approach, we try to mimic this cognitive process (Huang and Chang, 2023), using the human knowledge contained in LLMs and rationales as a means to produce the same associations human perform (Dasgupta et al., 2022; Strachan et al., 2024).

Tracking word senses and comparing them is a complex task (Kilgarriff, 1997). To address this problem, we approach it as a “comparative semantics” problem, i.e., instead of extracting the meaning as a final objective, we rely on the idea of relatedness and likeness of meaning to compute LSC. Similar to the work of Schlechtweg et al. (2024), we compare word occurrences for the characterization problem. However, in DUREL the annotator is prompted with two sentences that share a particular word and the annotator has to classify the level of similarity between contextual senses. In our approach, we use an LLM instead of humans to annotate the relatedness between words, and we reduce it to just ‘identical’ and ‘different’ classes.

We perform the characterization starting with a cognitive semasiological comparative analysis of the word meaning (Kilgarriff, 1997), following the Cambridge setting (Tang et al., 2013). We first provide a context where the word sense could be inferred (the Gracian approach (Agirre and Edmonds, 2007)), to then decide via reasoning if the deduced senses are different to a class of change (Blank, 2003).

This step is done by a LLM that acts like a judge/annotator using a special type of CoT (Wei et al., 2022) with detailed step-by-step reasoning (Mitra et al., 2023) to elicit models ability of word sense induction and comparative semantics. Our approach exploits rhetorical techniques to produce ‘cognitive-appealing’ arguments on how the senses are different.

For the first pole, **dimension**, we created a prompt requesting a word sense differentiation, where the prompt ask if a word is used in an identical or different sense. To perform the sense differentiation, we instruct the model to use **zeugma** (Kearns, 2006) as a cognitive approach to identify identical senses. If it can produce a consistent zeugma, the senses are identical. Otherwise, it should assume the words are different.

Zeugma is a rhetorical device where a single word, typically a verb or an adjective, governs or modifies two or more words in a sentence. This device creates a clever or unexpected relationship between different sentence parts. Zeugma often results in a play on words contributing to the overall impact of the expression, and it adds a layer of complexity or humor to the language used in a sentence and allows us to explore the sense usage difference (Kearns, 2006). For example:

(1) "He lost his keys."
(2) "He lost his temper."
(3) "He lost his keys and his temper." (?)

In sentence (3), the word “lost” is used to combine both sentences in a related sense to describe both (1) losing physical objects (keys) and (2) losing emotional control (temper). This zeugma creates a figurative and compact expression that links two different related meanings of the word ‘lost’ in a single sentence, creating a bad pun (Kearns, 2006). This bad pun comes because the second usage of ’lost’ does not preserve the same sense as the first usage, indicating a difference in the meaning. In Figure 2, we present part of the prompt employing zeugma for the dimension dataset. In Appendix D, we share the complete prompt for the experiments.

Sense Differentiation

[...]. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Write a sentence that joins both using zeugma and the same shared word while preserving the same sense. If the construction makes a bad pun, the words have a different sense.
- Step 4. Based on the previous reasoning, give your final answer: ’identical’ or ’different.’

[Few-shot examples.]

Figure 2: Prompt for sense differentiation in the dimension dataset.

This work proposes a computational instruction for figurative language analysis. The instruction is based on a simplified version of the Metaphor Identification Procedure (MIP) (Steen et al., 2007). To distinguish between metaphor and metonymy, the model is tasked with building a relation between the concepts. This relation can be either abstract, suggesting a metaphorical mapping between domains (evoking tropes), or material, if a physical association exists between the concepts.

(1) "The main objective of this forthcoming decision will be preparation for the winter."
(2) "Winter can cause many disruptions for public transport."
(3) "The word 'winter' in the second sentence is associated with its problems, such as snow, making it a metonymy."

The provided examples showcase how **simile** act as a parsing mechanism for the AI model. By leveraging similes, the model can reframe figurative language based on the similarity or association it expresses between concepts. In essence, a simile acts as a rhetorical device that explicitly compares two entities to enhance explanation and detail the nature of that comparison. By deciphering the figurative meaning within context, we aim to guide the model towards extracting more information about the underlying semantic relationship. This, in turn, allows the model to make a more accurate prediction regarding the type of relation – whether it's a metaphor or metonymy. In Figure 3, we illustrate the prompt for obtaining the figurative association between word usages.

Sense Figurativeness

[...]. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Compare the usage, determining if the second is related as a metaphor (where the word is used in a similar but non-literal sense), as a metonymy (where the word represents something closely related to or associated with it), or unrelated, used with a different sense.
- Step 4. Based on the third step, write the final answer, 'metaphor', 'metonymy', or 'unrelated.'

[Few-shot examples.]

Figure 3: Prompt for figurative sense in the relation dataset.

For **orientation** pole, the current state-of-the-art sense-level sentiment analysis requires first a WSD step, then a sentiment analysis step (Zhang et al., 2023). Similar to the previous prompts and following the best practice, we instruct the rationale to perform a textual sense disambiguation and then differentiate the orientations between these senses (Wiebe and Mihalcea, 2006). To differentiate orientation, we use **antanagoge** to compare senses' positiveness (or negativeness) and to enrich contextual information on how these senses can be perceived in the training data.

Antanagoge is a rhetorical device that involves responding to an accusation or negative point with a counter-argument or positive point. It is used to mitigate the impact of something negative by placing it alongside something positive. We use the common sentence "I'd rather X than Y" as a few-shot demonstration to instruct LLM to get the most probable contextual ordering. In the example below exemplify the usage of antanagoge.

(1) "A terrific presentation."
(2) "A terrific storm."
(3) "I'd rather have an terrific presentation than an terrific storm."

Sense Orientation

[...]. Follow this instructions to execute the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Leverage the rhetorical strategy of antanagoge, contrasting a negative with a positive, to weigh why one meaning might be more favorable than the other, or if they stand neutral.
- Step 4. Based on the third step, write the final answer 'negative', 'positive', or 'neutral.'

[Few-shot examples.]

Figure 4: Prompt for sense orientation in the orientation dataset.

Further optimization can be done as in (Schlechtweg et al., 2024). A graph of occurrences can be build and clusterized to extract senses over time. These clusters allows annotating semantic changes for each sense over time.

4 Experiments

In this section, we introduce the dataset produced to evaluate our annotation method and measure the quality of our annotations for LSC.

4.1 Dataset for Lexical Semantic Change Characterization

From a semasiological perspective, words' meaning could be inferred from the context, for example, "He *targeted* me, after I didn't agree with his proposal." or "The *mustache* guy, is coming today?". We can deduce the meaning of a word based on context and/or knowledge of the original meaning.

Lexical and Semantic Change (LSC) reflects how word meaning evolves. New senses emerge when a word is used in a novel, non-standard way. Over time, if this usage becomes widespread enough, it transitions from creative expression to a conventional meaning.

The ideal approach to LSC detection should mimic this human capability. This means employing unsupervised learning techniques, where the system infers the evolving sense of a word solely based on its prior exposure to the language and the contextual information within the data. In essence, the system learns to identify semantic shifts without the need for pre-labeled data (Schlechtweg et al., 2020).

To evaluate our framework, we produce an LSC Characterization dataset following the Cambridge setting described in (Tang et al., 2013). The dataset is composed of pairs of sentences sharing the same word (see the example with the word 'lost' from equation (1) and (2)). The first sentence expresses one possible usage (e.g., original usage), while the second sentence express a different usage. The task is to infer how the word's meaning in each context and compare them. We create three new datasets, one for each pole of change, where the instance pairs present the type of change the LLM should characterize in the pipeline.

In the **dimension** dataset, we curated the WiC data (Pilehvar and Camacho-Collados, 2019) for getting a fraction of reliable and high-quality examples. The original dataset only classified the word's meaning as related or unrelated. We adapted it according to the DUREL format. For this, the word's meanings are identical if the same meaning is observed when we merge the sentences (see the 'lost' example, equation (3)), i.e., a zeugma can be performed between the two sentences, relabeling the sentences to this.

We define words as related if they have a direct relation (metaphor or metonymy) for their usage, if the relation is not direct, we define this senses as unrelated. For example, if 'head' is used figuratively in both usages, the 'head' to represent the leader of a company and the 'head' to represent the mind, we define them as unrelated as there is no direct figurative usage between them. Finally, if not one of the cases above, we set them as unrelated or keep the original annotation.

For the **relation** data, we collected examples from the metaphor detection dataset (Choi et al., 2021) to get literal and metaphorical usages and also examples from the literature to increase the evaluation dataset size, the sentences were manually collected using online dictionaries like Linguee[1] and Merriam-Webster[2] and verified by 3 human annotators. To collect metonymies, we similarly used examples in literature (Lakoff and Johnson, 2008) and retrieved sentences from online sources.

The **orientation** data we created by getting sense pairs for the same word where we had the greatest variance from these pairs by analyzing SentiWordNet (Baccianella et al., 2010). The sentence pairs were obtained from SemCor (Raganato et al., 2017) and WordNet (Miller, 1995) depending on how easy it is to infer the sense given the sentence. Additionally, we transform the sentences so that the sentiment of a word cannot be trivially detected from the whole sentence, so the detector needs to comprehend the word-level sentiment. In some cases, we modified the sentence to be negative while the word meaning is positive.

Task	labels	Total
Dimension	Identical , Different	260
Relation	Metaphor, Metonymy, Unrelated	331
Orientation	Positive, Negative , Neutral	262

Table 2: General view of the three datasets created for Semantic Change Characterization.

In Table 2 we describe the number of instance pairs for each dataset we produced.

4.2 Experimental results

We compare our approach with two baselines to evaluate how good LLMs and rhetoric devices are for characterizing semantic change. The prompt is based on CoT and a few-shot prompt with no CoT, where all prompts are provided with the same 3-shot examples with the correct label and also the dictionary sense. We took a special care to the difference in prompt be only the method and

[1] www.linguee.com
[2] www.merriam-webster.com

not inserting ‘hack’ phrases[3] to improve model performance.

We selected LLaMA-3 and Phi-3 as the current state-of-the-art LLMs for the instruction prompt. We sampled the models 5 times for each method with temperature $\tau = 0.7$, using the guidance[4] library to control the generated layout. We report the mean and standard deviation of the accuracy.

In Table 3, we present the results for the dimension dataset. We can observe that the rhetoric method meaningfully improves the accuracy of Phi-3 and LLaMA-3-70b over the baselines, while for LLaMA-3-8b, the best method is few-shot prompt. While the data used for instruction tuning LLaMA-3 is not publicly released, we believe it was fine-tuned in WiC data (Pilehvar and Camacho-Collados, 2019), which could explain the improved accuracy.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
Few-Shot	**.75±.00**	.76±.00	.60±.00
CoT	.59±.02	.75±.02	.62±.01
Rhetoric (ours)	.68±.03	**.78±.01**	**.71±.02**

Table 3: Accuracy metrics for the Dimension dataset.

Table 4 shows our results for the relation dataset. For both models, the rhetoric method improved significantly over the baselines. With LLaMA-3-70b getting the overall best capabilities for recognizing figurative usage.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
Few-Shot	.38±.00	.52±.00	.49±.00
CoT	.45±.01	.53±.01	.51±.03
Rhetoric (ours)	**.52±.02**	**.59±.02**	**.56±.01**

Table 4: Accuracy metrics for the Relation dataset.

In Table 5, the accuracy is tied between few-shot and rhetoric for LLaMA-3-8b, while for Phi-3 is 10 points better. This task is particularly more difficult than the rest for LLMs.

Method	LLaMA-3-8b	LLaMA-3-70b	Phi-3
Few-Shot	**.37±.00**	.36±.00	.30±.00
CoT	.34±.03	.35±.00	.30±.02
Rhetoric (ours)	**.37±.03**	**.39±.01**	**.40±.02**

Table 5: Accuracy metrics for the Orientation dataset.

In Appendix C we provide a correlation plot across models and prompting strategies, we can observe that the Rhetorics method generate the lowest correlation between strategies. This evidence points out that the initial bias of the model for inferring change in meaning is very different from the current established methods in linguistic literature.

5 Discussion

In the boxes below, we present some cherry-picked cases of the reasoning approach with rhetoric devices. This illustrates how the method can improve over no reasoning.

Example #54 cup

- He drank a **cup** of coffee.
- He borrowed a **cup** of sugar.

In sentence 1, it is used with the sense of a container. — In sentence 2, it is used with the sense of an amount. — "He drank a cup of sugar." Both sentences use ’cup’ with different meanings.

A: different

In the example above the CoT and rhetorics method inference ‘cup’ in the second sentence is used for “an amount”, instead of “a container”. However, the rhetorics method differs by generating a zeugma that helps make clear why the senses don’t work, as drinking a cup of sugar sounds absurd.

Example #9 battle

- During the course of this **battle**, four Canadian soldiers were wounded.
- I call on him not to let the **battle** of hierarchies within the directorates-general prevent this from being implemented.

In the first sentence, ’battle’ is used literally, referring to a fight or conflict. — In the second sentence, ’battle’ is used metaphorically to represent a struggle or conflict between different groups or factions. — The word ’battle’ is related to the first sentence as a metaphor.

A: metaphor

For the figurative prompt, the Phi-3 model enriched the prompt with a reasonable inference for the second sentence, with this information

[3] e.g. "Please think carefully", "Take a deep breath", "This is important for my career."

[4] https://github.com/guidance-ai/guidance/

the model could correctly relate the senses as a metaphorical variation.

> **Example #167 salt**
>
> - She was **salty** because of what happened yesterday.
> - She **salts** her lectures with jokes.
>
> In sentence 1, 'salty' is used to describe someone who is angry or upset. — In sentence 2, 'salts' is used to describe the act of adding flavor to something, in this case, a lecture. — I'd rather have a lecture with jokes than be angry or upset. The second sense is more positive.
>
> A: positive

The orientation prompt forces an order between the senses, making comparing non-trivial perceptions of good and bad feasible. In the example above, the model is instructed to use antanagoge and then impersonate a preference over both scenarios; this allows it to correctly classify the second sense as more positive.

The illustrative examples demonstrate that LLMs can retrieve stored cultural knowledge and enrich prompts, allowing in-context learning to utilize more information for decision-making. However, in some cases, they fail to correctly manipulate senses to produce zeugma, and sometimes the conclusions are inconsistent with the reasoning. We have detailed these failure cases in Appendix A.

This observation aligns with well-documented phenomena: LLMs can hallucinate and generate incorrect reasoning even if they arrive at the correct answer (Ye and Durrett, 2022). Other failure cases may involve the leakage of evaluation data and the confusion of generalization with memorization, as it is difficult to verify whether an LLM-generated figurative usage explanation is novel or directly derived from training data.

Meaning is a fundamental open question in NLP. While LLMs can often replicate human-like behavior by relying heavily on form, they struggle with simpler tasks that require basic understanding of meaning (Berglund et al., 2023). Understanding how models deal with meaning in controlled settings (such as comparing a word with itself) is crucial for enabling models to generalize beyond mere form.

While rhetorical devices are standard tools in linguistics, our understanding of their cognitive and psychological effectiveness is still evolving. By using LLMs, we can better test theories and explore how humans relate meanings through language, further advancing our understanding in this area.

6 Limitations

Our sentences were carefully selected to enable a concrete understanding of the word's meaning in the context; if the method is applied in under-contextualized sentences, it can result in wrong sense disambiguation, leading to a bad overall performance. We evaluated our approach on LLMs optimized for English; while the methodology applies to other languages, the quality of characterization may differ. LLMs can hallucinate on generating explanations, so the generated rationale may not reflect the correct cognitive likeness between senses even when it correctly characterizes the change.

7 Conclusion

In this paper, we exploited the vast amount of cultural knowledge in LLMs to relate and characterize the word's meaning change for different contexts. Hereafter, we experimentally observed that rhetoric devices can help LLMs to mimic human meaning associations, enabling semantic change characterization.

LLMs has many rhetoric examples in its training data, which are not trivially detected. In future works, we expect to explore in depth if LLMs can produce new rhetoric associations or if it's just reproducing its training data, this task can be investigate through non conventionalized metaphor or circumstantial metonymy.

Another important aspect to investigate is how to better prompt language model for figurative reasoning if it can perform better by providing all forms of metaphors (instrumentality, container, etc.) (Lakoff and Johnson, 2008) and metonymies (part-whole, cause-effect, etc.) (Lakoff and Johnson, 2008) as examples. We also expect to extend the DUREL approach to automatize the characterization of semantic change.

References

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Keith Allan. 2013. *The Oxford handbook of the history of linguistics*. OUP Oxford.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *ArXiv*, abs/2309.12288.

Andreas Blank. 2003. Words and concepts in time: Towards diachronic cognitive onomasiology. *Trends In Linguistics Studies And Monographs*, 143:37–66.

Vladimir V. Bochkarev, Stanislav V. Khristoforov, Anna V. Shevlyakova, and Valery D. Solovyev. 2022. Neural network algorithm for detection of new word meanings denoting named entities. *IEEE Access*, 10:68499–68512.

Jader Camboim de Sá, Dimitra Anastasiou, Marcos Da Silveira, and Cédric Pruski. 2024. Socio-cultural adapted chatbots: Harnessing knowledge graphs and large language models for enhanced context awarenes. In *Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024)*, pages 21–27, St Julians, Malta. Association for Computational Linguistics.

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

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *ArXiv*, abs/2311.09022.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Ishita Dasgupta, Andrew Kyle Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *ArXiv*, abs/2207.07051.

Lauren Fonteyn and Enrique Manjavacas. 2021. Adjusting scope: A computational approach to case-driven research on semantic change. In *Workshop on Computational Humanities Research*.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. *ArXiv*, abs/2101.07668.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Päivi Juvonen and Maria Koptjevskaja-Tamm. 2016. *The lexical typology of semantic shifts*, volume 58. Walter de Gruyter GmbH & Co KG.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? Yes it is true because it weakens superficial cues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180, Dubrovnik, Croatia. Association for Computational Linguistics.

Kate Kearns. 2006. *The Handbook of English Linguistics*. Wiley Online Library.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Chenyang Lyu, Yongxin Zhou, and Tianbo Ji. 2022. MLLabs-LIG at TempoWiC 2022: A generative approach for examining temporal meaning shift. In *Proceedings of the First Workshop on Ever Evolving NLP (EvoNLP)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *ArXiv*, abs/2311.11045.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements. *ArXiv*, abs/2404.18570.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin UK.

Steven Pinker and Paula Jorde Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences*, 13:707 – 727.

Lidia Pivovarova and Andrey Kutuzov. 2021. Rushifteval: a shared task on semantic shift detection for russian. *ArXiv*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

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

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.

Gerard Steen, Lynne Cameron, Alan Cienki, Peter Crisp, Alice Deignan, Raymond W. Gibbs, Joe Grady, Zoltán Kövecses, Graham David Low, and Elena Semino. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S A Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature human behaviour*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv: Computation and Language*.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. Semantic change computation: A successive approach. *World Wide Web*, 19:375–415.

Elizabeth Closs Traugott. 2017. Semantic change.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *ArXiv*, abs/2312.06002.

Taylor W. Webb, Keith J. Holyoak, and Hongjing Lu. 2022. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526 – 1541.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Neural Information Processing Systems*.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish. *ArXiv*, abs/2205.06691.

Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Conference on Empirical Methods in Natural Language Processing*.

A Negative Examples

In this section, we present some examples where the model failed to produce the correct form or to give an answer over the correctly produced reasoning.

A.1 Dimension

Below are some examples of dimension data:

Example #44 come

- **Come** out of the closet!
- He **came** singing down the road.

In sentence 1 it is used with the sense of emerge. — In sentence 2 it is used with the sense of move in a certain direction. — "He came out of the closet singing." The word 'come' doesn't make a bad pun, but it has different senses. It is still different.

A: different

In the example above, the zeugma was not correctly constructed, and it ignored the instruction, which only makes sense if it makes a bad pun.

Example #67 do

- All you ever **do** is surf the Internet.
- A 'B' grade **doesn't** suffice to get me into medical school.

In sentence 1 it is used with the sense of an action. — In sentence 2 is used with the sense of to suffice, to be enough. — "A' B' grade does not suffice for me to do what I want." The word 'do' makes a bad pun, so is different.

A: different

In the example above, it failed to build a zeugma using both contexts, even though the answer is correct.

A.2 Orientation

Below are some examples of orientation data.

Example #67 do

- He do not resent they supervision or person 's sometimes tiring **sympathy**.
- The group exhibit more concern over the possibility of the abolitionist make a martyr of person than it do over the development of **sympathy** for he.

In the first sentence, 'sympathy' is used in a negative context, indicating annoyance or irritation. — In the second sentence, 'sympathy' is used to express understanding or concern. — Expressing understanding or concern (second sentence) is more positive than feeling annoyed or irritated.

A: positive

While the answer was correct, in the example above, the model didn't use antithesis for ordering senses.

B Inference Settings

Given hardware constraints, we used the gguf version of the models with `llama.cpp`[5] library. In the `Phi-3` (mini version), we used the 4k context with fp16 quantization. `LLaMA-3-8b` we used 8bit quantization and `LLaMA-3-70b`, 2bit quantization. All model weights were obtained from HuggingFace[6].

We did all the experiments on a Tesla V100 with 32GB RAM. The inference for all the models and prompts took less than four days.

C Correlation Across Prompt Strategies and Models

In this section, we present the correlation plot between judgments across different models and different prompts.

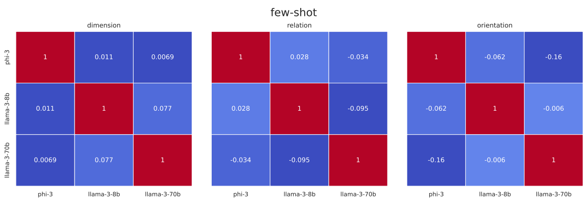

Figure 5: Correlation for Few-shot prompting.

In Figure 10 we observe that few-shot and CoT approaches are highly correlated.

[5] https://github.com/ggerganov/llama.cpp
[6] https://huggingface.co/models

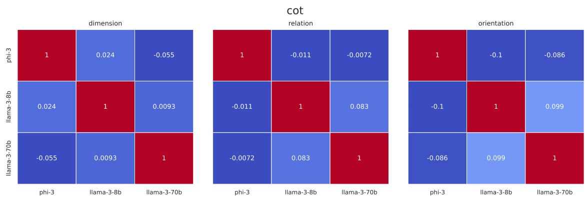

Figure 6: Correlation for CoT prompting across models.

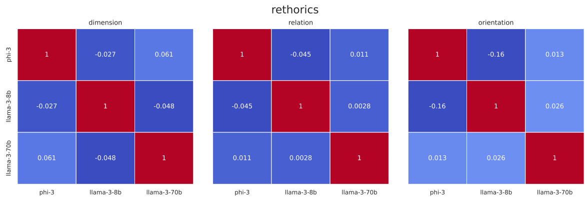

Figure 7: Correlation for Rhetorics prompting across models.

D Detailed Prompts

In the figures below, we show the detailed prompt for each type of characterization.

E Annotation

To obtain sentiment labels for the orientation data we relied on human annotation from volunteer students from a University in Europe where students have different backgrounds and different native language but English is used as the main language in their studies. The annotations were anonymously collected.

We first provided the annotators with the agreement terms: "This is a study on sentiment perception of polysemous words. This data will be freely available for research purposes. Inside you'll be asked to rate how the feeling varies for a word in different sentences. Your answers will be completely anonymous. COMPANY will not collect your personal data through this questionnaire and will not be able to identify you based on your answers. For more information about COMPANY's privacy notice please visit our webpage at: URL"

Then we presented a training screen in Figure 14.

We prompted the annotators for sentiment analysis with screen Figure 15.

F Ai Assistants In Research Or Writing

As our native language is not English, we used AI assistants like Grammarly, ChatGPT, and Gemini to improve vocabulary, grammar, and readability of this documents and the prompts. We also checked all generated text for inconsistencies with the original intent and fixed them properly when identified.

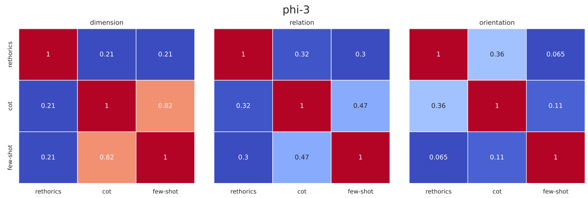

Figure 8: Correlation for Phi-3 model across strategies.

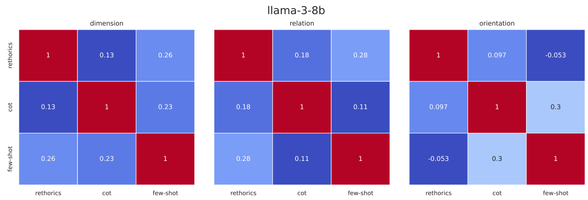

Figure 9: Correlation for LlaMA-3-8b model across strategies.

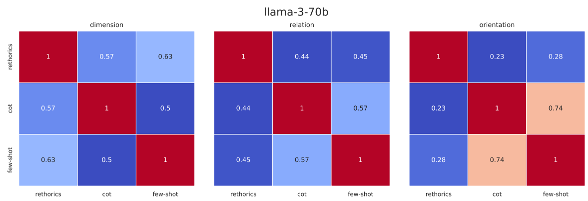

Figure 10: Correlation for LlaMA-3-70b model across strategies.

Sense Differentiation

You are presented with two sentences that both contain a specific word. Your task is to analyze how this word is used in each sentence and determine if its usage in the second sentence represents the same sense with respect to its use in the first sentence. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Write a sentence that joins both sentences using zeugma and the same shared word while preserving the same sense. If the construction make a bad pun, the words have different sense.
- Step 4. Based on the previous reasoning give your final answer: 'identical' or 'different.'

[Few-shot examples.]

Figure 11: Prompt for sense differentiation in the dimension dataset. The model is instructed to perform a zeugma association between senses to reason if it has a identical or different sense

Sense Figurativeness

You are presented with two sentences that both contain a specific word. Your task is to analyze how this word is used in each sentence and determine if its usage in the second sentence represents a metaphor or a metonymy with respect to its use in the first sentence. Follow these steps to complete the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Compare the uses, determining if the second is related as a metaphor (where the word is used in a similar but non-literal sense), as a metonymy (where the word represents something closely related to or associated with it), or unrelated, used with a different sense.
- Step 4. Based on the third reasoning, write the final answer, ’metaphor’, ’metonymy’, or ’unrelated.’

[Few-shot examples.]

Figure 12: Prompt for figurative sense in the relation dataset. The model is instructed to relate the meanings by association or similarity.

Sense Orientation

You will be provided with two sentences that share a common word used with different senses. Your task is to describe if the second sense for the word is more positive then the first. Follow this instructions to execute the task:

- Step 1. Describe the meaning of the word in the first sentence.
- Step 2. Describe the meaning of the word in the second sentence.
- Step 3. Leverage the rhetorical strategy of antithesis, contrasting a negative with a positive, to weigh why one meaning might be more favorable than the other, or if they stand neutral.
- Step 4. Based on the third reasoning, write the final answer ’negative’, ’positive’, or ’neutral.’

[Few-shot examples.]

Figure 13: Prompt for sense orientation in the orientation dataset. By using antithesis the model should order the senses polarity using ‘personal preference’ argumentation.

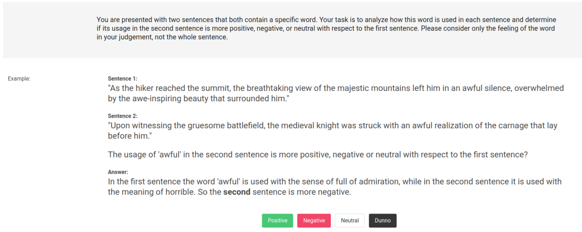

Figure 14: Training screen

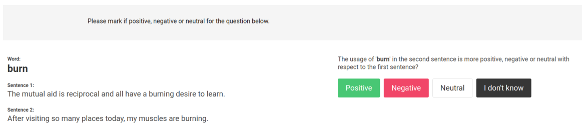

Figure 15: Annotation screen for sentiments.