SemBench: A Universal Semantic Framework for LLM Evaluation

Mikel Zubillaga, Naiara Perez, Oscar Sainz, German Rigau
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{*name.surname*}@ehu.eus

Abstract

Recent progress in Natural Language Processing (NLP) has been driven by the emergence of Large Language Models (LLMs), which exhibit remarkable generative and reasoning capabilities. However, despite their success, evaluating the true semantic understanding of these models remains a persistent challenge. Traditional benchmarks such as Word-in-Context (WiC) effectively probe this capability, but their creation is resource-intensive and often limited to high-resource languages. In this paper, we introduce SemBench, a framework for automatically generating synthetic benchmarks that assess the semantic competence of LLMs using only dictionary sense definitions and a sentence encoder. This approach eliminates the need for curated example sentences, making it both scalable and language-independent. We evaluate SemBench in three languages (English, Spanish, and Basque) spanning different levels of linguistic resources, and across a wide range of LLMs. Our results show that rankings derived from SemBench strongly correlate with those obtained from standard WiC datasets. Furthermore, our analysis demonstrates that only a small number of examples is required to achieve stable and meaningful rankings. Overall, SemBench provides a lightweight, adaptable, and data-efficient framework for cross-lingual evaluation of semantic understanding in LLMs.

Keywords: Evaluation Methodologies, Semantics, Word Sense Disambiguation

1. Introduction

In recent years, the field of Natural Language Processing (NLP) has experienced significant advancements, largely driven by the development of Large Language Models (LLMs). Trained on massive datasets, these models have demonstrated impressive capabilities in generating coherent, human-like text. Today, LLMs are not only employed for traditional tasks such as summarization and translation, but are also increasingly used as autonomous agents, programming assistants, and even literature reviewers (Jiang et al., 2025; Dong et al., 2025; Liao et al., 2024). This shift in how LLMs are applied has also prompted a reevaluation of how they are assessed—moving away from static benchmarks toward more dynamic, context-sensitive evaluations, often inspired by methodologies from reinforcement learning. However, most of the traditional NLP tasks remain unsolved, even for the best performing state-of-the-art approaches.

One such evaluation approach is the Word-in-Context (WiC) challenge (Pilehvar and Camacho-Collados, 2019). As the name suggests, WiC assesses a model's ability to distinguish between different senses of the same word based on context. Specifically, the task presents a target word used in two separate sentences and asks the model to determine whether the word carries the same meaning in both instances or reflects different senses. While this may appear straightforward–especially for *language* models–it has proven to be quite challenging, with performance often only slightly better than random guessing (Hayashi, 2025).

Developing a WiC dataset for a given language can be as simple as extracting sense-specific examples from a dictionary. However, many dictionaries either lack usage examples altogether or are restricted by licensing constraints. Alternatively, manually constructing a WiC dataset is resource-intensive, requiring significant effort from linguistic experts (Goworek et al., 2025).

In this work, we introduce **SemBench**, a novel fully automatic framework for evaluating the semantic competence of LLMs. Rather than relying on pre-constructed datasets, SemBench performs evaluation through generation, using only a dictionary with sense definitions (a resource typically more accessible than dictionaries containing usage examples) and a sentence encoder. This design makes SemBench both scalable and language-independent, enabling consistent evaluation even in low-resource settings. We apply our methodology across three typologically diverse languages (English, Spanish, and Basque) and a broad range of LLM families and sizes. Experimental results show that the model rankings by SemBench strongly correlate with those obtained from standard WiC datasets, validating its effectiveness. Moreover, ablation studies confirm the efficacy and practicality of SemBench, as only a small number of instances are required to achieve stable and meaningful results.

In summary, our main contributions are as follows: **(1)** we present SemBench, a fully automatic methodology for evaluating semantic understanding in LLMs through text generation, which yields results strongly aligned with WiC; **(2)** we demonstrate the adaptability of SemBench across languages

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4597–4610
11–16 May 2026.

with varying resources levels—high (English), moderate (Spanish), and low (Basque); **(3)** we analyze the impact of the number of test instances and benchmark size, showing that minimal data are sufficient to produce stable and interpretable rankings; and, **(4)** we propose a simple yet effective heuristic for controlling evaluation difficulty, which accurately reflects task complexity while preserving high correlation with WiC performance.

2. Related Work

2.1. Evaluating the semantic capabilities on LLMs

Although human evaluation remains the most reliable method for assessing the quality of generated responses, particularly since the emergence of LLMs, it is both time-consuming and difficult to scale when comparing multiple models or variations. With the rapid progress in the field, the need for efficient automated evaluation methods has become increasingly evident.

One of the most influential automatic benchmarks before the emergence of LLMs was SuperGLUE (Wang et al., 2019), a language understanding benchmark that extends the original GLUE benchmark (Wang et al., 2018). SuperGLUE introduced a set of more challenging sub-tasks designed to test various aspects of language model performance. Among these tasks is Word in Context (WiC; Pilehvar and Camacho-Collados, 2019), which focuses on evaluating the semantic capabilities for a given language model. More precisely, WiC presents a scenario in which the same word appears in two different contexts. The model is then asked to determine whether the word has the same meaning in both sentences or, conversely, whether each occurrence conveys a different sense. Although WiC was originally designed to evaluate context-sensitive word embeddings rather than LLMs, recent work has shown that it also serves as an effective benchmark for assessing LLM semantic understanding (Hayashi, 2025). In this work, we leveraged WiC to generate the model rankings that we will consider as gold standard for our methodology, SemBench.

2.2. Sense representations in LLMs

Transformer architectures have fundamentally changed representation learning by producing context-sensitive embeddings for tokens rather than single, fixed vectors. Whereas static embeddings assign one position per word type (Pennington et al., 2014), Transformer-based models produce a family of context-dependent vectors, which change according to the context. This richer representational palette has practical value for the study of lexical semantics (Petersen and Potts, 2023): geometric relationships among contextual embeddings offer a powerful means of distinguishing nuances in word meaning.

With the advent of recently developed LLMs, automatically generated sense definitions provide a new avenue for representing word usage (Periti et al., 2024). These definitions effectively capture the underlying meaning of a word, even in cases of polysemy, making them a valuable and interpretable tool for word usage representation (Gardner et al., 2022). In this work, we aim to indirectly measure the semantic capabilities of LLMs by measuring the quality of the generated definitions.

2.3. Generating Word Sense Definitions and Examples

Early works on generating word sense definitions were motivated by the goal of improving the interpretability of static embeddings (Gadetsky et al., 2018; Mickus et al., 2022). The original formulation enhanced models to generate a natural-language definition from a single static embedding of the target word (Noraset et al., 2017). Due to the polysemous nature of some words, producing accurate definitions from the word embedding alone proved difficult; consequently, the paradigm shifted: instead of relying solely on the word, contextual information was incorporated to generate more appropriate and precise definitions (Ishiwatari et al., 2019; Huang et al., 2021; Zhang et al., 2022). To do so, recent works have started using LLMs to create those definitions (Giulianelli et al., 2023; Periti et al., 2024).

In addition to definitions, context or example sentences where the word is used can also be generated. This is a key challenge in understanding and modeling language semantics. Some approaches rely on training models with existing dictionaries or corpora, pairing headwords with illustrative sentences to automatically create new examples (Barba et al., 2021; He and Yiu, 2022). In contrast, Harvill et al. (2023) introduced a more flexible method, showing that meaningful sentences can also be produced using just a single reference sentence as input. More recently, Cai et al. (2024) show that example generation can be done using LLMs in a zero-shot setting, only using as input the definition of the headword. However, previous works focus on definition generation without the purpose of evaluating model's semantic capabilities. In this work, we focus on this topic and present SemBench, an automatic framework for semantic capabilities evaluation.

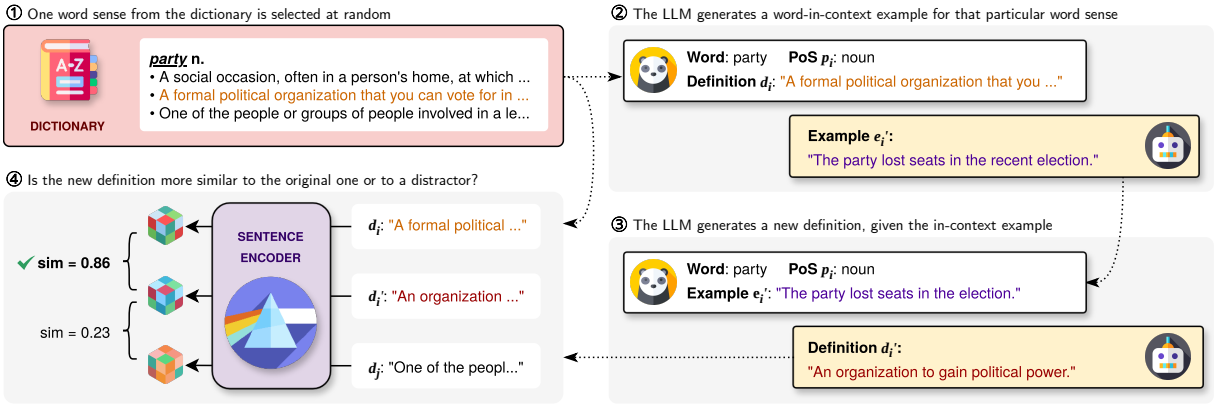


Figure 1: A general overview of the SEMBENCH framework.

3. SEMBENCH

SEMBENCH is a novel, fully automatic framework for evaluating the semantic understanding capabilities of LLMs. The approach rests on the intuition that a model demonstrating genuine semantic competence should be able to transition consistently between *definitions* and *usage examples* corresponding to the same sense of a word. Unlike previous evaluation protocols that rely on curated datasets or manual annotation, SEMBENCH constructs test instances directly from existing lexical resources. Its overall workflow is summarized in Figure 1, which illustrates how words, senses, and model generations interact within the framework. The remainder of this section details the core resources and methodological components of SEMBENCH.

3.1. Resources

SEMBENCH relies on two key components: a **sentence encoder** that is used to compute the semantic similarity of definition pairs, and a **dictionary**. Formally, let the dictionary be represented as

$$D = \{(w, S(w)) \mid w \in W\},$$

where W is the set of all words contained in the dictionary, and each word $w \in W$ is associated with a finite set of senses:

$$S(w) = \{s_1, s_2, \ldots, s_{n_w}\},\ n_w \geq 1.$$

Each sense $s_i \in S(w)$ is represented as a triplet $s_i = (d_i, p_i, e_i)$, where d_i denotes the textual definition of the sense, p_i is the part-of-speech (PoS) label, and e_i is an optional usage example. That is, every word is associated with one or more senses, each defined by at least a definition and grammatical category and, possibly, accompanied by an in-context example.

3.2. Methodology

SEMBENCH begins by sampling a polysemous word $w \in W$, such that $n_w > 1$. From its set of senses $S(w)$, one particular sense s_i is chosen at random and used as the seed for generation. Depending on the available dictionary format, SEMBENCH provides two experimental configurations:

- **From definitions** (SEMBENCH$_{\text{Def}}$): the default setup, shown in Figure 1, which does not assume access to in-context examples. The LLM is asked to *i)* generate a usage example e'_i for w given its definition d_i and PoS p_i; and *ii)* generate a dictionary definition d'_i for word w given its PoS p_i and the synthetic example e'_i.

- **From examples** (SEMBENCH$_{\text{Ex}}$): a simpler setup that assumes the dictionary provides a usage example e_i. The LLM is asked to generate a dictionary definition d'_i for word w given its PoS p_i and the context e_i (that is, it bypasses step ② in Figure 1).

The synthetic definition d'_i is then compared against two reference definitions from the same dictionary entry: the **target definition** d_i, which corresponds to the intended sense, and a **distractor definition** d_j, associated with a different sense s_j of the same word w ($i \neq j$). The model is considered correct if its definition is more semantically similar to the target than the distractor, according to an encoder-based similarity metric:

$$sim(d'_i, d_i) > sim(d'_i, d_j),$$

where similarity is computed as the dot product of the corresponding embedding representations. Model performance is then quantified as the proportion of correctly identified senses over a set of N randomly selected test instances (d_i, d_j).

Language	Reference WiC	Dictionary	Sentence Encoder
(EN) English	Pilehvar and Camacho-Collados (2019)	Oxford Dictionary of English (ODE)[1]	EmbeddingGemma 300M (Vera et al., 2025)
(ES) Spanish	Vázquez Abuín and Garcia (2025)	*Diccionario de la RAE* (DRAE)[2]	
(EU) Basque	Urbizu et al. (2022)	*Egungo Euskararen Hiztegia* (EEH)[3]	

Table 1: Resources employed in the experimentation for each language.

4. Experimental Setup

To validate the effectiveness of SemBench, we compare its results against a well-established word sense disambiguation evaluation framework: Word-in-Context (WiC). Specifically, we evaluate a set of LLMs with both methodologies and measure the degree of correlation between the two frameworks using the **Spearman's rank correlation coefficient** (ρ). A high correlation would indicate that SemBench captures a notion of semantic understanding comparable to traditional sense discrimination tasks, while requiring no manual annotation.

Moreover, we conduct experiments in three typologically diverse languages with varying levels of resource coverage, each of which has an existing WiC benchmark for reference: **English** (Germanic/high), **Spanish** (Romance/moderate), and **Basque** (isolate/low).

Below, we first detail the WiC datasets employed to establish reference rankings, then introduce the resources used to build and run SemBench in each language, and finally outline the evaluated LLMs and their inference setups.

4.1. Computation of Reference Rankings

Table 1 reports the existing WiC datasets that we rely on to establish a reference ranking of models per language. These datasets contain pairs of sentences featuring the same target word w used in two distinct contexts. Each pair is annotated with a binary label indicating whether w conveys the *same sense* in both contexts or not. To evaluate each LLM on this task, we follow a configuration similar to SemBench$_{\text{Ex}}$: for each WiC instance, the model is presented with the two contexts e_i and e_j in which w appears, and is asked to generate a dictionary-style definition for each occurrence, denoted d'_i and d'_j. Then, we encode both d'_i and d'_j using the same encoder employed throughout the experiments (introduced next), and compute their cosine similarity. The model is considered to predict *same sense* if $sim(d'_i, d'_j) > 0.5$,[4] and *different sense* otherwise. Accuracy is then calculated as the proportion of correctly classified WiC pair instances according to the gold labels.

Statistic	English	Spanish	Basque
Sense density	6.12 $\pm$ 3.5	6.92 $\pm$ 5.9	2.07 $\pm$ 0.3
Definition length	64.77 $\pm$ 27.7	59.77 $\pm$ 31.7	39.99 $\pm$ 28.1
Example length	106.10 $\pm$ 42.5	n/a	n/a

Table 2: Statistics of SemBench $_{\text{RAND}}$. Lengths are reported in terms of number of characters.

4.2. Experimental Resources

For each language, we created a **test set of 1,000 instances** sampled at random from the corresponding lexical resource (see Table 1; our code and processed resources are available online).[5] Each instance consists of a target definition and a distractor definition belonging to a different sense of the same word, selected according to the procedure described in Section 3.2.

SemBench$_{\text{Def}}$ vs SemBench$_{\text{Ex}}$. As shown in Table 2, only the English subset includes example sentences alongside sense definitions. Consequently, the comparisons between SemBench$_{\text{Def}}$ and SemBench$_{\text{Ex}}$ were conducted exclusively for English. To ensure a fair comparison, we used the same set of senses across both evaluation variants. The Spanish and Basque subsets were evaluated using SemBench$_{\text{Def}}$ only.

Difficulty levels. To better characterize the evaluation space, we constructed four sub-datasets of varying difficulty by controlling the semantic similarity between the target and distractor definitions. For each word, all alternative definitions were first ranked by their cosine similarity to the target definition, computed using the sentence encoder. Then, the distractor was selected according to one of four strategies:

- EASY: the least similar definition.
- MID: a definition from the middle of the list.
- HARD: the most similar definition.
- RAND: a randomly selected definition.

[1] www.oed.com; through Valera and Rigau (2021).

[2] www.rae.es/; accessed via the public website.

[3] www.ehu.eus/eeh; SQL dump 2024-01-26.

[4] Although the range of the similarity metric is defined as $[-1, 1]$, we considered 0.5 as the default decision threshold because 0 is interpreted as not similar and 1 is interpreted as perfectly similar.

[5] https://github.com/MikelZubi/SemBench

For each difficulty level, we sampled 1,000 pairs of senses from the English dictionary, in order to compare its effect on the SemBench$_{\text{Def}}$ and SemBench$_{\text{Ex}}$ strategies.

Sentence encoder. As for the sentence encoder, we chose the EmbeddingGemma model of 300M parameters (Vera et al., 2025) for all the experiments, given its multilingual support and demonstrated competitiveness.

4.3. Models and Inference Details

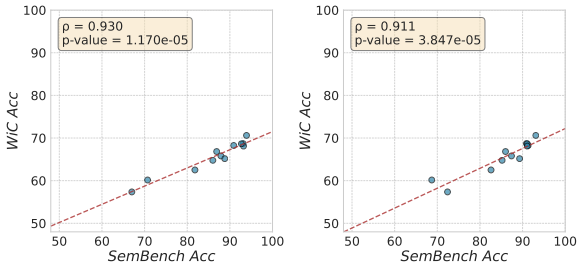
We evaluated a diverse set of open-weight LLMs, covering a range of architectures, sizes, and training paradigms. All models were accessed through their instruction-tuned or chat variants to ensure consistent prompting behavior. The evaluated models include:

- **Gemma 3** (Team et al., 2025): Multimodal instructed decoder only models. We have used the 4B, 12B, and 27B Instruct variants.
- **Qwen3** (Yang et al., 2025): Reasoning instructed decoder only models. We have used the 4B, 8B, 14B, and 32B Instruct variants.
- **Llama 2** (Touvron et al., 2023): Instructed decoder only models. We have used the 7B Instruct variant.
- **Llama 3.1** (Grattafiori et al., 2024): Instructed decoder only models. We have used the 8B and 70B Instruct variants.
- **Latxa Instruct** (Sainz et al., 2025): Instructed decoder only models specialized in the Basque language, built on Llama 3.1. We have used the 8B and 70B variants.

We ran experiments under both **zero-shot** and **5-shot** configurations, following the same prompting scheme across all languages and SemBench variants SemBench$_{\text{Def}}$ and SemBench$_{\text{Ex}}$. Prompts were designed to elicit definitions or examples as described in Section 3.2; full templates are provided in Appendix A. To ensure reproducibility, we generated all outputs using a greedy decoding strategy. Regarding the examples used for few-shot prompting, whenever possible we sampled definition–example pairs for each sense. However, since not all dictionaries include examples alongside definitions, we manually created the missing ones. As only five examples were needed, this did not require substantial manual effort.

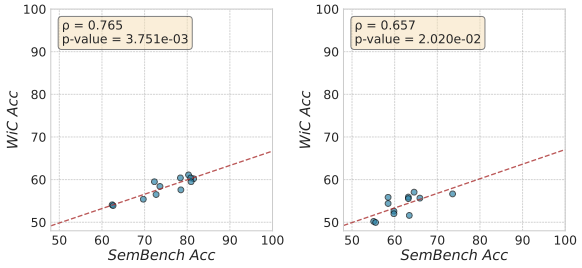
5. Results

To validate our hypothesis, we present the following experimental results. We begin by comparing the rankings produced by our proposed method with those obtained using WiC, our gold standard. Next, we extend the analysis to moderate- and low-resource language evaluations to demonstrate the applicability of SemBench. We also investigate how the number of examples in SemBench affects the final rankings, highlighting the robustness of our approach. Finally, we present the overall results obtained with SemBench.



(a) SemBench$_{\text{Def}}$ (b) SemBench$_{\text{Ex}}$

Figure 2: Accuracy (Acc) correlation between SemBench and English WiC using 5-shot configuration. On the left, we compare SemBench$_{\text{rand}}$ starting from definitions, whereas on the right, we compare the variant starting from examples.



(a) Spanish (b) Basque

Figure 3: Accuracy (Acc) correlation between SemBench$_{\text{Def}}$ with Spanish (left) and Basque (right) WiC datasets on the 5-shot settings.

5.1. Validation of SemBench against WiC

Our analysis covers three languages representing different levels of linguistic resources, allowing us to assess the robustness and general applicability of the method. The following sections present and discuss the results obtained for each language.

Validation against English WiC. Figure 2 illustrates the correlation between the scores obtained by the two variants of our proposed method (namely, SemBench$_{\text{Def}}$ and SemBench$_{\text{Ex}}$) and those from WiC for English. All the scores are obtained using 5 few-shot examples. We observe that, although our benchmarks tend to yield slightly higher absolute scores (up to 95 vs 70 points), they exhibit an almost perfect correlation with WiC ($\rho = 0.930$ and

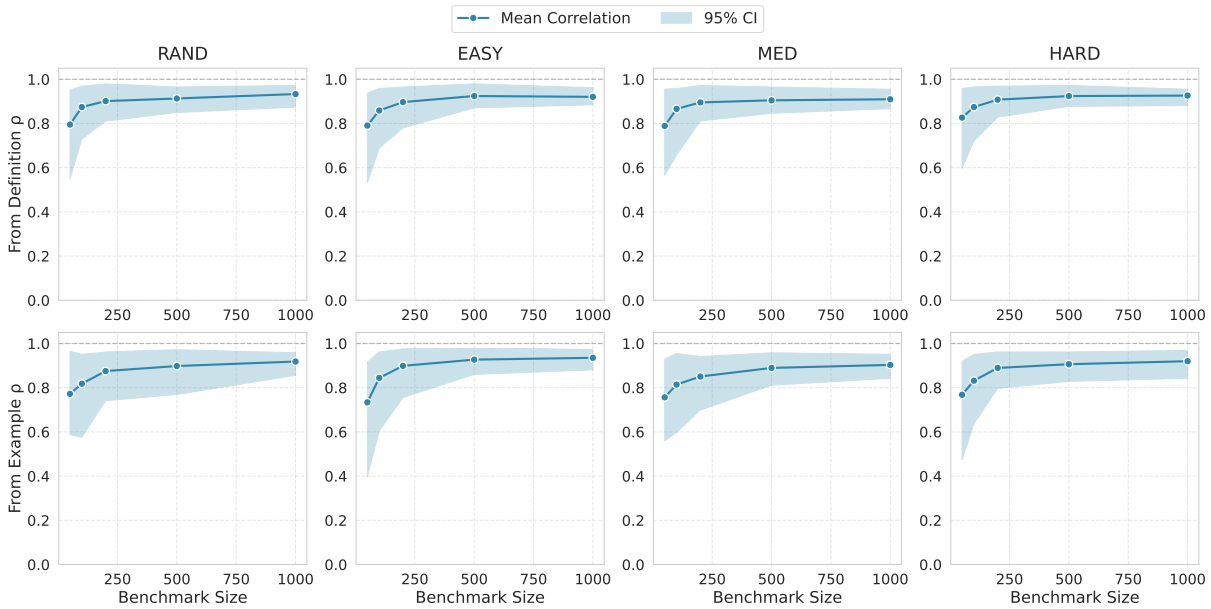


Figure 4: Spearman's ρ correlation curves between different SemBench variants and WiC with varying data-points. Confidence intervals are obtained by applying bootstrapping through 100 iterations.

$\rho = 0.911$), demonstrating the strong reliability and validity of our approach. Moreover, the range of values across models is noticeably broader, which suggest that SemBench offers greater discriminative capacity; in contrast, WiC results tend to cluster more tightly, particularly among high-performing models. A closer examination reveals only minor differences between the two variants, SemBench$_{\text{Def}}$ (Figure 2a) and SemBench$_{\text{Ex}}$ (Figure 2b). This observation further reinforces our hypothesis that a model with genuine semantic understanding should maintain consistent performance when transitioning between *definition*-based and *example*-based evaluations.

Validation against Spanish WiC. Following the same procedure as in English, we evaluated the results obtained by SemBench against the Spanish WiC dataset, as shown in Figure 3a. In this evaluation, we report only the results for the SemBench$_{\text{Def}}$ variant, since the Spanish dictionary used in our setup does not provide example sentences for every sense. The results yield a promising Spearman's correlation coefficient of $\rho = 0.765$, indicating that our method is also effective and reliable for moderately resourced languages such as Spanish. Furthermore, we confirm that SemBench provides higher discriminative power than WiC, with results showing again a noticeably wider spread across models in Spanish.

Validation against Basque WiC. Results for Basque are presented in Figure 3b. The Basque results exhibit a lower but still statistically significant rank correlation ($\rho = 0.657$, p-value < 0.05). This can be largely attributed to the fact that most models perform at near-random levels on the Basque WiC (as shown below in Section 5.3), leading to almost random rankings. Although this phenomenon also affects SemBench to some extent, its results appear more coherent: Basque-specialized models consistently outperform the others, indicating that SemBench is still capable of capturing meaningful distinctions even under low-resource conditions. Interestingly, SemBench remains more sensitive than WiC to relative performance differences, even when overall accuracy is low.

5.2. Robustness analysis

In addition to the validation with WiC on several languages, we also explored the robustness of our method regarding the total amount of examples in the benchmark required to produce a significantly well correlated ranking and the number of in-context examples (zero-shot vs few-shot). The following paragraphs discuss the results we obtained.

Impact of the number of instances. Figure 4 shows the Spearman's correlation curves as a function of the number of test instances. To conduct this analysis, we performed bootstrapping over 100 iterations on subsets randomly sampled from the original test set of 1,000 instances, with sample sizes ranging from 50 to 1,000 instances. This procedure was repeated for the two SemBench variants (SemBench$_{\text{Def}}$ and SemBench$_{\text{Ex}}$) across four difficulty levels: *random*, *easy*, *medium*, and *hard*.

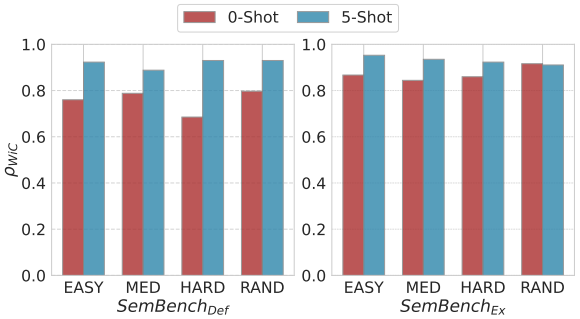


Figure 5: Spearman's ρ correlation comparison between 0-Shot and 5-Shot in SemBench$_{Def}$ (left) and SemBench$_{Ex}$ (right) against WiC.

For each point, we report the correlation along with the corresponding 95% confidence interval.

Overall, all variants converge rapidly to a high correlation (above 0.9), with only marginal gains observed beyond 500 instances. As expected, confidence intervals narrow as the number of data points increases, indicating greater stability with larger sample sizes. Both variants, *from definitions* and *from examples*, display comparable correlation curves; however, contrary to our expectations, SemBench$_{Def}$ yields consistently smaller confidence intervals. A plausible explanation for this behavior is that the examples generated by the models are more similar among them, leading to slightly more stable results for the SemBench$_{Ex}$ variant.

Zero-Shot vs. Few-Shot Results Comparison. All results reported so far were obtained by prompting the LLMs with five in-context examples. This few-shot configuration helps guide the model toward the desired task format and can improve its alignment with the evaluation objective. However, this setup also introduces a small amount of manual effort, since examples must be created when they are not available in the dictionary. Although generating five examples per sense represents only minimal human intervention, it raises an important question: to what extent do these examples influence the overall performance and model ranking? To address this, we compare the results obtained under two settings—zero-shot (no examples provided) and few-shot (five examples)—and report the difference in Spearman's ρ between the two.

Figure 5 summarizes these results across all experimental configurations. Overall, the correlations remain high in both scenarios, indicating that the models capture the intended task dynamics even without explicit examples. However, we observe decreases in correlation values when transitioning from few-shot to zero-shot prompting. The largest performance gap appears in the SemBench$_{Def}$ approach, where the removal of examples leads to a more noticeable drop in correlation. This behavior is expected, as SemBench$_{Def}$ involves additional intermediate steps in the evaluation pipeline. The model must first generate a usage example from the definition and then infer the definition again from that generated example, which introduces additional opportunities for variation. Consequently, the presence of in-context examples may provide useful guidance that stabilizes this multi-step generation process.

	SemBench$_{Ex}$			SemBench$_{Def}$		
Model Name	**EASY**	**MED**	**HARD**	**EASY**	**MED**	**HARD**
Gemma 3_{4B}	78.10	67.20	61.40	79.30	68.80	63.70
Qwen3_{4B}	**91.20**	**86.20**	**82.00**	**92.00**	**86.40**	**83.50**
Llama 2_{7B}	78.70	68.50	66.00	75.50	64.60	60.80
Llama 3.1_{8B}	90.80	84.00	81.60	91.20	84.90	82.20
Latxa$_{8B}$	87.60	81.10	77.00	86.00	77.90	74.60
Qwen3_{8B}	**93.10**	**87.80**	**86.10**	**95.00**	**90.10**	**86.70**
Gemma 3_{12B}	88.40	83.50	78.10	90.30	85.00	79.60
Qwen3_{14B}	**93.90**	**89.10**	**85.70**	**96.20**	**91.50**	**89.90**
Gemma 3_{27B}	89.70	83.80	80.00	91.00	84.60	80.40
Qwen3_{32B}	**95.30**	**89.40**	**86.70**	**96.90**	**92.90**	**91.30**
Llama 3.1_{70B}	**94.00**	**89.10**	85.50	95.50	90.50	**89.30**
Latxa$_{70B}$	93.90	88.80	**86.30**	**95.60**	**90.80**	87.80

Table 3: SemBench English results in 5-Shot scenario in the different difficulty levels. Bold-case indicates the best results by group. Underline indicates the best results overall.

The correlation remains remarkably stable in SemBench$_{Ex}$, demonstrating that the approach maintains its reliability and consistency even in a zero-shot setting. Overall, these results underscores the potential of SemBench as a scalable evaluation framework, minimizing the dependence on handcrafted examples while preserving strong agreement with human or reference-based judgments. In practical terms, these results suggest that SemBench can be effectively applied in scenarios where annotated data or examples are scarce, further reinforcing its suitability for low-resource or multilingual contexts.

5.3. SemBench results

Overall SemBench performance of all evaluated models across languages and difficulty levels is summarized in Tables 3 and 4. The former presents English results for both SemBench$_{Def}$ and SemBench$_{Ex}$ variants by difficulty level, while the latter extends the comparison to the multilingual setup (English, Spanish, and Basque) using the random difficulty strategy.

Impact of SemBench difficulty. We observe a consistent trend across all configurations: as task

Model Name	English		Spanish		Basque	
	SB	WiC	SB	WiC	SB	WiC
Gemma 3_{4B}	70.70	60.16	62.60	53.90	55.08	50.20
Qwen3$_{4B}$	**88.80**	**65.16**	**78.50**	**57.60**	**59.82**	**52.00**
Llama 2_{7B}	67.00	57.36	62.40	54.10	55.53	49.95
Llama 3.1_{8B}	87.90	65.80	72.70	56.50	59.82	52.60
Latxa$_{8B}$	81.80	62.51	69.70	55.40	**63.43**	51.60
Qwen3$_{8B}$	**90.90**	**68.25**	**80.80**	**60.40**	58.47	**54.40**
Gemma 3_{12B}	86.00	64.77	72.30	59.50	**63.21**	55.55
Qwen3$_{14B}$	**93.20**	**68.11**	**81.50**	**60.20**	**63.21**	**55.90**
Gemma 3_{27B}	86.90	66.83	73.60	58.40	58.47	55.85
Qwen3$_{32B}$	**93.90**	**70.61**	**80.30**	**61.10**	**64.56**	**57.05**
Llama 3.1_{70B}	**93.10**	**68.69**	78.40	**60.40**	65.91	55.65
Latxa$_{70B}$	92.70	68.65	**80.90**	59.50	**73.59**	**56.65**

Table 4: Multilingual results in WiC and SemBench$_{Ex}$ (SB) with RAND strategy at 5-shot scenario. Bold-case indicates the best results by group. Underline indicates the best results overall.

difficulty increases, model performance systematically decreases, following the pattern EASY > MED > HARD. This confirms that SemBench's difficulty control mechanism effectively captures the semantic complexity of the generated instances, making the use of difficulty-controlled subsets convenient when available (although random sampling also yields stable results). Similarly, both SemBench variants yield closely aligned results, with SemBench$_{Def}$ generally producing slightly higher scores than SemBench$_{Ex}$, suggesting that examples generated by LLMs provide richer contextual cues that can lead to more accurate sense discrimination when generating definitions.

Impact of model family and size. When examining specific model families, the Qwen3 series stands out as the best overall performer, surpassing even larger models in most configurations. This advantage likely stems from its enhanced reasoning-oriented training, which appears to contribute more precise sense discrimination and definition generation. Larger variants within each family (e.g., Llama 3.1 70B and Qwen3 32B) consistently outperform their smaller counterparts, capturing the positive effect of scaling on semantic generalization. Conversely, smaller models such as Gemma 4B and Llama 7B show pronounced drops in accuracy as difficulty increases. Interestingly, linguistically specialized models (Latxa 8B and 70B) demonstrate that language adaptation can partially offset size limitations, especially in low-resource settings.

Impact of SemBench language. The multilingual results in Table 4 provide further insight into performance trends across resource availability levels. As expected, absolute performance decreases from English to Spanish and further to Basque, mirroring the availability of linguistic resources and, potentially, the quality of underlying dictionaries. Still, SemBench successfully preserves meaningful model rankings even in low-resource settings, with Basque-specialized models (namely, Latxa) outperforming general-purpose ones in the Basque test—unlike in WiC. That is, SemBench can capture language-specific semantic competence effectively, even when absolute accuracy is limited.

6. Conclusions

We introduce SemBench, a framework for evaluating the semantic competence of Large Language Models (LLMs) through controlled text generation from dictionary definitions. By relying solely on sense definitions and a sentence encoder, SemBench provides a fully automatic, language-independent approach that eliminates the need for manually annotated data.

Our experiments confirm that SemBench produces model rankings strongly aligned with those obtained on standard WiC benchmarks, validating it as an effective and faithful alternative to semantic evaluation. At the same time, the wider range of results observed in SemBench indicated a higher discriminative capacity, enabling to separate model performances more clearly and reveal subtler differences in semantic competence. Multilingual evaluations further demonstrate that SemBench maintains consistency across English, Spanish, and Basque. Particularly in Basque, domain-adapted models outperform general-purpose ones, unlike in WiC, which shows that SemBench can better capture language-specific semantic competence even under limited-resource conditions.

Ablation studies also confirm the scalability and efficiency of the approach: as few as 250 instances are sufficient to produce interpretable results, and only marginal gains are observed beyond 500. In addition, the proposed difficulty-control heuristic accurately reflects task complexity, with performance decreasing predictably from EASY to HARD sets while preserving strong correlation with WiC.

Overall, SemBench offers a consistent, interpretable, and language-independent methodology for assessing semantic competence in LLMs. Its robustness across configurations, model sizes, and languages confirm its value as a lightweight and convenient alternative to traditional manually annotated benchmarks. Most importantly, by depending only on dictionary definitions and general-purpose encoders, SemBench can be readily applied to new and under-resourced languages where benchmarks such as WiC or other annotated datasets are not available.

Limitations

While SemBench shows strong alignment with traditional semantic benchmarks, several limitations remain. First, SemBench relies on a strong multilingual encoder (i.e., EmbeddingGemma 300M), which has demonstrated competitive performance. However, our dependence on a single encoder may introduce biases and may fail to distinguish between subtle senses. Exploring alternative or ensemble encoders could further enhance robustness. Second, our evaluation focused on open-weight models of different sizes. However, including commercial LLMs such as Claude or GPT would help assess whether SemBench can effectively rank state-of-the-art language models. Finally, it would be interesting to compare SemBench results with more general evaluation efforts, such as LLMArena, to examine whether the semantic competence measured here aligns with broader model preferences and overall perceived quality.

Acknowledgments

This work has been partially supported by the European Union under Horizon Europe (Project LUMINOUS, grant number 101135724) and the Spanish Ministry of Science, Innovation, and Universities (Project HumanAIze, grant number AIA2025-163322-C61). It was also funded by the Basque Government (IKER-GAITU project) and the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project Desarrollo de Modelos ALIA. Mikel Zubillaga holds a PhD grant from the University of the Basque Country UPV/EHU (PIF24/04).

References

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3779–3785. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Bill Cai, Ng Clarence, Daniel Liang, and Shelvia Hotama. 2024. Low-cost generation and evaluation of dictionary example sentences. 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 3538–3549, Mexico City, Mexico. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Jiaru Qian, Tian Wang, Kechi Zhang, Zhi Jin, and Ge Li. 2025. A survey on code generation with LLM-based agents.

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

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: literature review and dataset analysis. *Applied Computing and Intelligence*, 2:83–98.

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.

Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve,

Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt,

Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models.

John Harvill, Mark Hasegawa-Johnson, Hee Suk Yoon, Chang D. Yoo, and Eunseop Yoon. 2023. One-shot exemplification modeling via latent sense representations. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 303–314, Toronto, Canada. Association for Computational Linguistics.

Yoshihiko Hayashi. 2025. Evaluating LLMs' capability to identify lexical semantic equivalence: Probing with the word-in-context task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6985–6998, Abu Dhabi, UAE. Association for Computational Linguistics.

Xingwei He and Siu Ming Yiu. 2022. Controllable dictionary example generation: Generating example sentences for specific targeted audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

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

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

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. 2025. Towards rationality in language and multimodal agents: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3656–3675, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as research tools: A large scale survey of researchers' usage and perceptions.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and

word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

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

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically generated definitions and their utility for modeling word meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for Basque. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila

Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

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

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Leire Valera and German Rigau. 2021. Adquisición de conocimiento léxico a partir de diccionarios.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and lightweight text representations.

Marta Vázquez Abuín and Marcos Garcia. 2025. Assessing lexical ambiguity resolution in language models with new WiC datasets in Galician and Spanish. *Procesamiento del Lenguaje Natural*, 74(0):305–319.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang

Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022. Fine-grained contrastive learning for definition generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1001–1012, Online only. Association for Computational Linguistics.

A. Prompts

In order to define the tasks of SemBench for LLMs, we designed two complementary prompt configurations: *i)* generating an example from a definition, and *ii)* generating a definition from an example. In line with best practices for prompting modern LLMs, each configuration consists of a *System Prompt* and a *User Prompt*.

A.1. Generating an example from a definition

System Prompt: You are an expert *{language}* lexicographer. Your task is to generate ONLY ONE example in *{language}* of the usage of a word, given a definition of that word. Please, provide JUST the example—DO NOT include the definition or any other further explanation.

User Prompt: Given the *{part-of-speech}* '*{word}*' and its sense in this definition: '*{definition}*', generate one usage example of the word for that sense. Give JUST the example without further explanation.

A.2. Generating a definition from an example

System Prompt: You are an expert *{language}* lexicographer. Your task is to generate a dictionary definition in *{language}* of a word, given some example sentences of the word. Please, provide JUST the definition—DO NOT include the example or any other further explanations.

User Prompt: Given the *{part-of-speech}* '*{word}*' and its sense in this example: '*{example}*', generate the definition of the word for that sense. Give JUST the definition without further explanation.