Are Large Language Models Good Word Sense Disambiguators?

Anonymous ACL submission

Abstract

Word Sense Disambiguation (WSD) is a historical task in computational linguistics that has received much attention over the years. However, with the advent of Large Language Models (LLMs), interest in this task (in its classical definition) has decreased. In this study, we evaluate the performance of various LLMs on the WSD task. We extend a previous benchmark (XL-WSD) to re-design two subtasks suitable for LLM: 1) given a word in a sentence, the LLM must generate the correct definition; 2) given a word in a sentence and a set of predefined meanings, the LLM must select the correct one. The extended benchmark is built using the XL-WSD and BebelNet. The results indicate that LLMs perform well in zero-shot learning but cannot surpass current state-of-the-art methods. However, a fine-tuned model with a medium number of parameters outperforms all other models, including the state-of-the-art.

1 Introduction

The Word Sense Disambiguation (WSD) task (Ide and Véronis, 1998; Navigli, 2009; Bevilacqua et al., 2021) has a long tradition in computational linguistics. The most used definition of WSD is selecting the correct meaning for a word occurrence in a text from a set of possible meanings provided by a sense inventory. The classical definition requires the existence of a sense inventory that provides for each word a list of possible meanings. This inventory can be a dictionary, a thesaurus, or a semantic network such as WordNet (Miller, 1995). However, since the existence of a sense inventory can be a limit in particular contexts, the task of Word Sense Discrimination/Induction was introduced. In this case, the task is to infer the different word usages by clustering the word occurrences according to their meaning; in this way, occurrences that share the same meaning are grouped in the same cluster, and the final set of clusters corresponds to the differing meanings of that word.

In the past, several techniques were exploited to solve both tasks. In particular, the WSD methodologies evolved according to advances in Artificial Intelligence and Machine Learning. The first period was characterized by using rules-based systems followed by knowledge-based approaches when digital sense inventory became available. With the availability of digital corpora, supervised approaches were introduced to take advantage of manually annotated data. With the advent of the web, large corpora and large knowledge graphs automatically extracted for the web revolutionized supervised and knowledge-based approaches. A new set of approaches was proposed when language models based on the transformers' architecture (Vaswani, 2017) were introduced. The ability of these models to represent words in context through dense vectors opens new possibilities for the disambiguation and discrimination of word meanings.

The more recent novelty that revolutionized computational linguistics is the introduction of Large Language Models (LLMs). Essentially, an LLM is based on the transformer architecture trained on vast amounts of text data to understand and generate human-like language. LLMs have proven their ability to solve different tasks in a zero-shot or few-shot setting without using specific training data. However, it is also possible to fine-tune an LLM on specific tasks using training data. The capability of LLMs to solve several tasks without training suggests an intrinsic ability to understand the semantics behind the language. The impressive results achieved by LLMs might make us lose sight of or underestimate the problem of automatic disambiguation of meaning. In this work, we want to measure how state-of-the-art LLMs can solve the WSD task to understand if the model somehow stores knowledge about word meanings. The WSD task must be redesigned considering the generative abilities of LLMs. Therefore, we extend a previous benchmark with two subtasks suitable for testing

an LLM. We re-design the WSD task in two ways: 1) the model is tested in generating the definition of a word in a sentence; 2) the model is evaluated in selecting the correct meaning of a word in a sentence from a predefined set of possible choices following a multiple choices paradigm often used to evaluate LLMs.

Our study considers only **open** LLMs with a different number of parameters and several languages: English, Spanish, French, Italian and German. The performance is evaluated according to a gold standard, considering the quality of the generated definition (sub-task 1) and the correctness of the selected sense (sub-task 2).

The main contributions of our work are 1) the extension of an existing multilingual benchmark for testing and training LLMs in the context of the WSD task; 2) an extensive evaluation of open state-of-the-art LLMs; 3) the release of several fine-tuned models trained on our dataset[1].

The paper is structured as follows: Section 2 discusses related works that leverage LLMs for solving WSD; Section 3 provides details about the benchmark used to evaluate LLMs in the WSD task, while Section 4 describes the methodology used to generate and select answers exploiting LLMs. Results are discussed in Section 5, and final remarks are reported in Section 6.

2 Related Work

Transformer-based language models are widely used for solving the WSD task. A deep overview is proposed in (Loureiro et al., 2021). BERT (Devlin, 2018) and its variations excel in understanding context-sensitive semantic nuances, making them dominant in evaluation benchmarks. The authors find that BERT-like models can accurately distinguish between different word senses, even with limited examples for each sense. The analysis also shows that while language models can nearly solve coarse-grained noun disambiguation under ideal conditions (ample training data and resources), such scenarios are rare in real-world applications, leaving significant challenges. Moreover, the article compares two WSD strategies: fine-tuning and feature extraction. The authors conclude that feature extraction is more robust, especially in dealing with sense bias and when training data is limited. Notably, averaging contextualized embeddings as a feature extraction method is effective, performing well even with just few training sentences per word sense. However, no works about the usage of recent LLMs for WSD are proposed. In our work, we try to investigate the ability of LLMs in a zero-shot setting without any training data and considering a fine-grained sense inventory. Moreover, we also propose an analysis of a fine-tuned LLM when training data are available.

Another interesting analysis of WSD approaches based on BERT-like models is proposed in (Bevilacqua et al., 2021). This work analyzes several WSD approaches including ones that leverage language-models both for extracting contextual-embeddings used as features and as starting point for training a supervised model on sense-annotated data. This paper is strongly related to our work because it provides an extensive evaluation of the same dataset we considered.

Another interesting work is the one proposed by (Cabiddu et al., 2023), where several language models, including large models such as GPT and GPT-2, are evaluated in three behavioural experiments used to measure children's sense disambiguation capabilities. The study is interesting because it tries to compare how semantics is perceived by children and how it is represented in transformer-based models. The authors find a model bias with respect to the most dominant meaning and a negative correlation between the training size and the model performance. However, the authors limit their analysis to this dataset, which is very specific.

We find a unique work that uses LLMs on a task similar to WSD. In (Kritharoula et al., 2023), the authors combine transformer-based methods for multimodal retrieval and LLMs to solve the task of Visual WSD (VWSD). VWSD is a novel task that aims to retrieve an image among a set of candidates that better represents the meaning of an ambiguous word within a given context. LLMs are used as knowledge bases to enhance the textual context and resolve ambiguity related to the target word.

In conclusion, several works exploited transformer-based architectures similar to BERT for WSD essentially in two ways: 1) extraction of contextual embeddings used as features; 2) supervised models that fine-tune the language model on sense-annotated data. However, no works have exploited recent decoder-only LLMs as word sense disambiguators in a zero-shot setting (completely unsupervised) or as a base for further fine-tuning on annotated data. Our work tries to

[1]We plan to release code, models, models' outputs and datasets in case of acceptance.

fill this gap by considering the more recent and state-of-the-art open LLMs.

3 The Benchmark

To evaluate LLMs on the WSD task, we need a sense-annotated corpus, i.e., a collection of sentences in which each word is tagged with its correct meaning taken from a sense inventory. For this reason, we also require a sense inventory that provides the set of possible meanings for each word. Therefore, our benchmark requires both a multilingual corpus and a multilingual sense inventory.

We will introduce some formal notations before delving into the description of the benchmark construction. Given a sentence S_k and one of its word occurrences w_i, we define L_i as the list of possible meanings of w_i and $m_j \in L_i$, the meaning assigned to w_i. Each meaning has several glosses, one for each language taken into account, and we use $m_{j,lang} \in L_i$ to refer to it. In our case, $lang \in \{en, it, es, fr, de\}$. Starting from the multilingual sense-annotated corpus and the corresponding sense inventory, we need a strategy for building two types of prompts for testing LLMs.

The first prompt aims to assess the ability of the LLM to generate an accurate definition of a word within a specific sentence. For each sense annotated word occurrence, we create the prompt (Liu et al., 2023) in Table 1 for each language. The table reports only the prompt for English; the others are provided in the Appendix B. We also store the correct definition m_j in the benchmark in a field called `output`.

Prompt template (generation)
Give a brief definition of the word "w_i" in the sentence given as input. Generate only the definition. Input: "S_k"
English prompt
Give a brief definition of the word "art" in the sentence given as input. Generate only the definition. Input: "The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world."

Table 1: Prompt for the generation benchmark.

While constructing the prompt, we need to manage the cases in which a word w_i occurs more than once in the sentence S_k. In these cases, we change the prompt as follows: "*Give a brief definition of the x occurrence of the word "w_i"...*", where $X = \{first, second, third, fourth, fifth\}$ and $x \in X$. We exclude cases where the word occurs more than six times, and we translate the set X according to each language.

The goal of the second kind of prompt is to evaluate the LLM's ability to select the correct sense from a set of predefined possibilities following a multiple-choice paradigm. In this case, we exploit the list of all possible meanings L_i. In particular, from L_i, we remove all the annotated meanings[2] and obtain the set C_i. Then, we randomly add to C_i one of the correct meanings; in this way, C_i contains only one correct sense. For each occurrence of a sense-annotated word in the corpus, we create the prompt in Table 2 for each language. The table reports only the prompt for English; the others are provided in the Appendix B. Additionally, we store the identifier (i.e. the option's number) corresponding to the correct answer in a field called `output`.

Prompt template (multiple choice)
Given the word "w_i" in the input sentence, choose the correct meaning from the following: C_i. Generate only the number of the selected option.
English prompt
Given the word "art" in the input sentence, choose the correct meaning from the following: 1) Photographs or other visual representations in a printed publication 2) A superior skill that you can learn by study and practice and observation 3) The products of human creativity; works of art collectively 4) The creation of beautiful or significant things. Generate only the number of the selected option. Input: "The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world."

Table 2: Prompt for the multiple choice benchmark.

We also manage the case where the word w_i occurs more than once by modifying the prompt

[2] In the sense-annotated corpus, a word occurrence can be annotated with more than one correct meaning.

as in the first benchmark. Moreover, given that the model is asked to choose among different options in this benchmark, we need to manage cases in which the size of C_i is less than two. In these cases, we remove the occurrence from the dataset. Monosemic words are not considered in the construction of both tasks[3].

We use XL-WSD (Pasini et al., 2021) as our sense-annotated corpus. This dataset serves as a cross-lingual evaluation benchmark for the WSD task, featuring sense-annotated development and test sets in 18 languages from six different linguistic families. Additionally, it includes language-specific training data, making it highly useful for evaluating WSD performance in a multilingual context. As stated previously, this study is focused on five languages: English, Italian, Spanish, French, and German. The sense inventory adopted in XL-WSD is BabelNet (Navigli and Ponzetto, 2010). However, not all senses in BabelNet have a gloss for each of the chosen languages. For this reason, we build two versions of the dataset: **without translation** in which we consider only the word occurrences that have glosses in BabelNet for each language, and **with translation** in which English glosses[4] are automatically translated when they are not available in BabelNet for a particular language. We use the 1.3B variant of the Meta NLLB-200 model[5] for the translation. We selected this translation model because it has a good performance and computational cost trade-off. Moreover, it is open.

The Table 3 reports the number of instances for each kind of task. Statistics for each language are reported in Appendix A.

Without translation		
	Generation	Multiple-choice
Training set	1,204,430	861,791
Test set	10,480	9,847
With translation		
	Generation	Multiple-choice
Training set	1,451,650	1,170,921
Test set	11,473	11,168

Table 3: Task statistics: number of instances.

[3]For the first benchmark based on definition generation, it is also possible to consider monosemic words. We exclude this hypothesis since we want to test LLMs in the case of polysemy.

[4]The English gloss is always available.

[5]https://huggingface.co/facebook/nllb-200-1.3B

4 Methodology

We follow two distinct methodologies to evaluate LLMs in solving the WSD task. In the first approach, known as (**zero-shot**), we directly prompt a selection of open LLMs with a varying number of parameters, without any task-specific training, to assess their inherent ability to solve the WSD problem. In the second approach, we **fine-tune** an open LLM with a small number of parameters. Our aim is twofold: 1) testing existing open models in solving the disambiguation task without additional training and 2) determining whether a model with a small number of parameters can solve the disambiguation task with proper fine-tuning. We select a model with a small number of parameters to allow the training on more accessible hardware.

4.1 Zero-shot

The methodology of this approach is straightforward. We select a set of open LLMs with a different number of parameters and directly prompt them using the benchmarks described in Section 3. Then, we measure the quality of the generated definitions and the accuracy in selecting the correct sense from the predefined alternatives. We perform two separate evaluations: the former involves only the original glosses, while the latter also contains the machine-translated ones.

For prompting the models, we use a cloud service[6]. We consider the following LLMs: Llama-3.1-instruct-8B (Dubey et al., 2024), Mistral-instruct-7B-v03 (Jiang et al., 2023), Gemma2-9B (Team and el., 2024), Llama-3.1-instruct-70B, Qwen2-72B-Instruct (Yang et al., 2024) and Llama-3.1-instruct-405B.

All models are tested using a greedy search approach as an inference strategy to avoid variability over different runs.

4.2 Fine-tuning

We use LLAMA 3.1 8B INSTRUCT as the base model to fine-tune. The LLAMA 3.1 family of models has been designed and trained with multilinguality in mind, by properly balancing the languages in the training mixture. Therefore, the LLAMA 3.1 models are already skilled in understanding and generating text in multiple languages. Specifically, they support the following

[6]together.ai: https://www.together.ai/. We spend about 15$ for performing all the experiments.

languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. All languages considered in this work are also natively supported by LLAMA 3.1 8B INSTRUCT, making it the ideal starting point.

We use a full-parameter training approach for the fine-tuning strategy, using *DeepSpeed ZeRO 3*[7] (Rajbhandari et al., 2020) for parallelization. We rely on a compute node consisting of four A100 64GB VRAM GPUs. We use the LLAMA 3.1 instruct template to format the prompt without any system message. We use a maximum sequence length of 512, discarding examples exceeding this value. The value was selected after studying the number of tokens in the training sets with and without translation (also considering the tokens added by the instruction formatting). In all cases, the 95th percentile of the number of tokens was less than 512. Other relevant hyperparameters are reported in Table 4.

Parameter	Value
batch size	512
lr	4e-5
lr schedule	cosine
lr warmup ratio	0.00
weight decay	0
epochs	1
optimizer	AdamW

Table 4: Hyperparameters for fine-tuning

5 Evaluation

This section reports the results of both evaluations: the first involves several LLMs in a zero-shot setting, while the second is based on a fine-tuned model based on LLAMA 3.1 8B INSTRUCT.

We select two metrics to evaluate the quality of generated definitions: 1) RougeL refers to the overlap of longest co-occurring in sequence n-grams between the reference text and the generated one; 2) BERTscore (Zhang et al., 2019) exploits the pre-trained contextual embeddings from BERT and matches words in reference and generated definition by cosine similarity. RougeL gives us an idea of the syntactic similarity, while the BERTscore measures semantic coherence. The BERTscore is necessary since the LLM can generate a most extended or lexically different definition that is semantically correct. For the computation of the BERTscore we use an English BERT model for evaluating the English part of the dataset, while we exploit a multilingual version of BERT for the other languages. We use accuracy to measure the LLM's ability to select the correct sense from a set of alternatives.

We explore different settings since we use machine translation to create the missing glosses for all the languages. We perform zero-shot experiments on two benchmark subsets: 1) without machine translation and 2) with machine translation. We perform four runs when a fine-tuned model is involved, considering that translated glosses can also be used during the fine-tuning. Table 5 shows all the combinations and the corresponding labels used to reference each configuration. We perform only two runs for the zero-shot evaluation since we do not consider training data and need to test on two sub-sets: one with translation and one without translation.

Training data	Test data	Label
Without translation	Without translation	FF
Without translation	Machine translated	FT
Machine translated	Without translation	TF
Machine translated	Machine translated	TT

Table 5: Different evaluation settings according to the kind of glosses: original glosses in BabelNet and machine translation of missing glosses.

5.1 Zero-shot results

In this section, we report the zero-shot evaluation results. Table 6 shows results without machine translation, while Table 7 with translation. Gemma2-9B is the best medium model in the multiple-choice task (accuracy) for all languages without translation. We can observe similar performance between llama3.1-8B and Gemma2-9B in terms of the quality of generated definitions. Llama3.1-8B achieves the better RougeL score for all languages, while BERTscores are similar to Gemma2-9B. Mistral-7B achieves the worst results, but it is also the smallest model.

As expected, larger LLMs provide better results than the medium counterparts. Interestingly, Llama3.1-70B and Llama3.1-405B provide similar results despite the significant difference in the number of parameters. While Llama3.1-405B achieves the best accuracy for English, its performance is only slightly better than that of Llama3.1-70B. No-

[7] https://github.com/microsoft/DeepSpeed

	Llama3.1 8B-Instruct			Mistral 7B-Instruct			Gemma2 9B-Instruct		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
EN	**.2260**	.8638	.5587	.1149	.8314	.6171	.2116	**.8650**	**.6762**
IT	**.1363**	.6985	.4604	.0747	.6532	.5324	.1221	**.6986**	**.5840**
ES	**.1811**	**.7262**	.5802	.1408	.6872	.5898	.1570	.7158	**.6503**
FR	**.1901**	**.7247**	.5090	.1437	.6888	.6290	.1208	.6815	**.6493**
DE	**.1586**	**.7050**	.6217	.1101	.6808	.6130	.1091	.6791	**.6826**
	Llama 3.1 70B-Instruct			Qwen2-72B-Instruct			Llama 3.1 405B-Instruct		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
EN	**.2437**	.8654	.7520	.1670	.8455	.7370	.2393	**.8669**	**.7532**
IT	.1439	.7018	.6298	.1131	.6773	**.6396**	**.1524**	**.7072**	.6259
ES	.1900	.7231	.7012	.1749	.7096	**.7214**	**.1915**	**.7297**	**.7214**
FR	.1713	.7054	.7059	.1596	**.7057**	**.7624**	**.1751**	.7003	.7149
DE	**.1454**	**.6991**	.7739	.1276	.6955	.7652	.1343	.6894	**.7957**

Table 6: Zero-shot results without machine translation.

tably, Qwen2-72B-Instruct achieves good results for Spanish and French.

Results with machine translation are shown in Table 7. The results are pretty similar to those of the experiments without machine translation. Among medium-sized models, Gemma2-9B provides better accuracy, but Llama3.1-8B provides the best quality in the generation. Llama3.1-405B achieves the best accuracy and generation quality for larger models, although Llama3.1-70B occasionally provides similar or better performance. Only for Italian, the Qwen2-72B model has the best accuracy. These findings indicate that machine translation does not significantly affect the performance behaviour between LLMs. However, we observe a general decrease in performance across all LLMs and metrics.

5.2 Fine-tuning results

This section reports results for the Llama3.1-8B fine-tuned model (Llama3.1-8B-FT). We consider different training sizes (10K, 20K and the whole dataset (ALL)) and different data subsets with or without translated glosses during the training and testing, following the configurations shown in Table 5. For 10K and 20K subsets, we maintain a balanced distribution between the generation and the multiple-choice tasks. Specifically, we randomly select x instances for each language for each task type, where x corresponds to either 10K or 20K. Therefore, the whole training dataset consists of 100K and 200K for the 10K and 20K filtering, respectively.

The FF and FT settings results for the models trained without machine translation are in Table 8. These results refer to the model fine-tuned only on the original glosses without machine translation. It is essential to highlight that results for English change only when the training set varies since the machine translation does not affect the English test set. Generally, accuracy increases with the size of the training, except for German. German has fewer training instances, and its performance is affected when the whole training set is used since our model is trained simultaneously in all languages.

Results of the model fine-tuned on machine-translated glosses are reported in Table 9. The impact of machine translation during the training is minimal. If we consider tables 8 and 9, where during the test, we do not use machine translation, we observe similar results. Some languages are more affected by the introduction of machine translation since these increase the number of instances, but not all languages are equally represented in the training data.

Introducing translated instances in the test set impacts performance for many reasons. First, the number of instances to test increases, and we may introduce instances with more polysemy, especially in the multiple-choice benchmark, in which we add new glosses and potentially introduce more alternatives.

To compare our results with the ones proposed in (Pasini et al., 2021), we must consider the test set with translated glosses to cover all the instances in the original dataset. We only consider configurations on the multiple-choice benchmark since the XL-WSD evaluation metric (F1) is based on sense prediction. The quality of the definition generation cannot be compared with other systems since the other WSD approaches do not generate a sense definition. Since the multiple-choice task is also performed as a generative problem, we need to transform the answer provided by the LLM into the BabelNet sense id used by the XL-WSD scoring tool. This process is not trivial and requires several

	Llama3.1 8B-Instruct			Mistral 7B-Instruct			Gemma2 9B-Instruct		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
EN	**.2260**	.8638	.5587	.1149	.8314	.6171	.2116	**.8650**	**.6762**
IT	**.1318**	.6934	.4054	.0741	.6513	.4619	.1227	**.6971**	**.5304**
ES	**.1740**	**.7231**	.4575	.1339	.6823	.4749	.1519	.7131	**.5142**
FR	**.1717**	**.6934**	.4942	.1299	.6807	.5506	.1129	.6819	**.6274**
DE	**.1512**	**.6934**	.5728	.0961	.6671	.5432	.1018	.6776	**.5975**
	Llama 3.1 70B-Instruct			Qwen2-72B-Instruct			Llama 3.1 405B-Instruct		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
EN	**.2437**	.8654	.7520	.1670	.8455	.7370	.2393	**.8669**	**.7532**
IT	.1416	.7000	.5688	.1145	.6747	**.5837**	**.1500**	**.7053**	.5716
ES	.1851	.7211	.5676	.1650	.7034	.5721	**.1858**	**.7265**	**.5766**
FR	.1554	**.7012**	**.6991**	.1430	.6934	.6940	**.1594**	.6952	.6799
DE	**.1263**	**.6928**	.6444	.1191	.6817	.6617	.1150	.6850	**.6642**

Table 7: Zero-shot results with machine translation.

	Llama3.1-8B-Instruct-FT / 10K			**Llama3.1-8B-Instruct-FT / 20K**			**Llama3.1-8B-Instruct-FT / ALL**		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
Configuration FF									
EN	.4584	.9021	.7788	.5346	.9139	.7889	.7392	.9466	.8067
IT	.4739	.8068	.7580	.5649	.8350	.7881	.7452	.8920	.8234
ES	.5166	.8362	.8156	.6098	.8680	.8329	.7649	.9191	.8694
FR	.6108	.8629	.8937	.6831	.8904	.9253	.7923	.9258	.9163
DE	.6484	.8659	.9043	.6904	.8802	.8870	.7106	.8890	.8739
Configuration FT									
EN	.4584	.9021	.7788	.5346	.9139	.7889	.7392	.9466	.8067
IT	.4139	.7856	.6648	.4920	.8100	.7032	.6436	.8581	.7499
ES	.4046	.7942	.6416	.4600	.8123	.6551	.5689	.8482	.6924
FR	.4484	.8026	.8156	.4913	.8189	.8323	.5581	.8405	.8464
DE	.4476	.7884	.8346	.4705	.7969	.8741	.4837	.8021	.8395

Table 8: Evaluation results of the fine-tuned model trained on data without machine translation.

	Llama3.1-8B-Instruct-FT / 10K			**Llama3.1-8B-Instruct-FT / 20K**			**Llama3.1-8B-Instruct-FT / ALL**		
	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy	RougeL	BERTscore	Accuracy
Configuration TF									
EN	.4586	.9018	.7776	.5493	.9163	.7877	.7446	.9477	.8224
IT	.4154	.7880	.7410	.5336	.8245	.7737	.7128	.8815	.8234
ES	.4175	.8051	.8204	.4981	.8308	.8367	.6972	.8982	.8655
FR	.5666	.8481	.9072	.6530	.8745	.8914	.7779	.9205	.9186
DE	.5678	.8373	.9043	.6669	.8721	.8478	.7079	.8860	.8652
Configuration TT									
EN	.4586	.9018	.7776	.5493	.9163	.7877	.7446	.9477	.8224
IT	.3825	.7767	.6621	.4920	.8096	.6967	.6703	.8662	.7575
ES	.3790	.7877	.6499	.4469	.8085	.6744	.6460	.8729	.7079
FR	.4623	.8100	.8233	.5290	.8301	.8105	.6497	.8716	.8399
DE	.4573	.7897	.8247	.5143	.8090	.8222	.5553	.8250	.7852

Table 9: Evaluation results of the fine-tuned model trained on data with machine translation.

steps:

1. We extract the choice from the LLM answer using a regular expression;
2. The choice is used to extract the gloss from the instruction provided to the LLM;
3. The gloss is used to retrieve the sense id by searching over the glosses of the possible senses of the target word;
4. The multiple-choice dataset does not contain instances of monosemic words. We select the only available sense id as the prediction in these cases.

Analyzing Tables 8 and 9, we observe that when the translated test set is used, English, Italian and Spanish achieve the best accuracy with the whole machine-translated training set. Conversely, French and German perform best with the training set without machine translation: French performs best with the whole training set, while German with 20K instances. We decide to consider the TT setting for all the languages since, in a real scenario, using different training for each language is not feasible.

	EN	IT	ES	FR	DE	AVG
XLMR-Large	.7628	.7766	.7585	.8388	.8318	.7937
XLMR-Base	.7450	.7673	.7655	.8233	.8213	.7845
BERT-L	.7677	-	-	-	-	-
BERT-M	-	.7616	.7466	.8164	.8063	-
LS-BERT	-	.7388	.7477	.8078	.8213	-
QInterf	.8010◇	.7980	**.7900**	.8500	.8500	-
Llama3.1-8B-FT / ALL	**.8652**	**.8205**	.7769	**.8836**	**.8898**	**.8472**
Gemma2-9B (0-shot)	.7343	.6295	.5997	.7370	.8016	.7004
Llama3.1-70B (0-shot)	.7988	.6646	.6424	.7888	.8237	.7437
Llama3.1-405B (0-shot)	.8040	.6699	.6509	.7759	.8329	.7467

Table 10: XL-WSD results. ◇ The QInterf system is evaluated on a different portion of the English dataset that does not include data from SemEval-2010.

In Table 10, we report the F1 computed using the official scoring tool released by the XL-WSD creator. The table also shows the results of current best systems: *XLMR-Large* and *XLMR-Base* based on (Conneau et al., 2020), supervised approaches based on BERT. Moreover, we report several BERT-based systems: BERT-L is the large model specific for English, BERT-M is the multilingual model and LS-BERT is the language-specific model. Finally, we add a recent system QInterf (Zhang et al., 2024) based on a quantum interference model that calculates the probability that the target word belongs to a superposition state representing the multiple glosses of the same word sense. Table 10 also shows the best LLMs in the zero-shot setting and our fine-tuned model (*Llama-8B-FT / ALL*). Our fine-tuned model performs best for all languages, with a remarkable result of **.8652** for English. The results allow some interesting considerations:

- LLMs in zero-shot learning are not able to overcome the baseline models except large models that provide better results for English and German;
- all LLMs in zero-shot setting show poor performance for Italian and Spanish;
- a medium model (Llama3.1-8B) fine-tuned on training data provides impressive results and always overcomes large models and baselines, resulting in state-of-the-art performance.

6 Conclusions and Future Work

In this work, we investigate the performance of several LLMs in solving the WSD task. For that purpose, we extend an existing benchmark (XL-WSD) to support two new subtasks: 1) given a word occurrence in a sentence, the LLM must provide the correct definition; 2) given a word occurrence in a sentence and a set of predefined meanings, the LLM must select the correct on. To build our benchmark, we exploit the XL-WSD dataset and BebelNet. Moreover, we use training data available in XL-WSD for fine-tuning and LLM based on Llama3.1-8B. Results show that LLMs can provide good performance in zero-shot learning but are not able to overcome current state-of-the-art approaches. The best performances are obtained by large models, while medium ones provide poor results. However, the fine-tuned model with a medium number of parameters is able to overcome all the models, including the current state-of-the-art approaches. The fine-tuned model can achieve an impressive accuracy of **.8472** averaging all languages, and a remarkable accuracy of **.8652** in English.

7 Limitations

The current version of our work presents some limitations summarized in the following points:

1. Not all the languages presented in XL-WSD are taken into account. We focused on languages with adequate resources to ensure a robust evaluation pipeline. However, we plan to extend the analysis to underrepresented languages.
2. The few-shot approach is not considered. We have decided to exclude this approach to reduce the complexity of the paper. For the same reason, we did not consider a multi-prompt evaluation. Nonetheless, we are aware of their potential and will explore them in the future.

3. The fine-tuning approach involves only one model with a medium number of parameters, excluding larger models used in the zero-shot evaluation. This choice was made to investigate strategies that can be implemented on affordable hardware.

4. While the exclusion of ChatGPT may be seen as a limitation, we aim to promote the use of open models to improve reproducibility and transparency in research. In line with this, we also use Llama3.1-405b, which provides performance similar to ChatGPT-4o in several state-of-the-art benchmarks.

8 Ethical considerations

Our work is heavily based on pre-trained LLMs developed by external organizations. The pre-training procedure was performed without our supervision, and the datasets used for pre-training and fine-tuning were also not checked. Therefore, the models may produce inaccurate or biased results that reflect the potential issues present in the original training data.

To reduce inaccuracies, human experts manually checked the prompt templates for each language. These experts participated voluntarily and were fully informed about our research objectives and the use of the data they checked, ensuring transparency in their involvement and contributions.

References

M Bevilacqua, T Pasini, A Raganato, R Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *IJCAI*, pages 4330–4338. International Joint Conferences on Artificial Intelligence.

Francesco Cabiddu, Mitja Nikolaus, and Abdellah Fourtassi. 2023. Comparing children and large language models in word sense disambiguation: Insights and challenges. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1):1–40.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multimodal retrieval for visual word sense disambiguation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13053–13077, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proc. of AAAI*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Gemma Team and el. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Junwei Zhang, Ruifang He, Fengyu Guo, and Chang Liu. 2024. Quantum interference model for semantic biases of glosses in word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19551–19559.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix A

This appendix reports some detailed statistics about the dataset. Table 11 shows the number of instances for each kind of benchmark and language without glosses translation. Meanwhile, Table 12 reports the same statistics when machine translation is exploited to create missing glosses.

Without machine translation

	generation		selection	
	training set	test set	training set	test set
EN	565,831	6,757	421,213	6,605
IT	242,343	1,673	192,962	1,529
ES	216,317	1,248	153,817	1,041
FR	121,014	539	63,429	442
DE	58,925	263	30,370	230

Table 11: Number of instances for each language without machine translation.

With machine translation

	generation		selection	
	training set	test set	training set	test set
EN	*565,831*	*6,757*	*421,213*	*6,605*
IT	308,903	1,888	271,864	1,823
ES	345,258	1,601	319,156	1,554
FR	157,787	812	113,307	781
DE	73,871	415	45,381	405

Table 12: Number of instances for each language with machine translation.

B Appendix B

This appendix reports prompts in languages other than English. Native-language speakers or translation experts have checked all prompts. Some sentences have grammatical issues since the XL-WSD dataset could contain data obtained from machine-translated corpora. Prompts for Italian are in tables 13 and 14. Prompts for Spanish are in tables 15 and 16. Prompts for French are in tables 17 and 18. Prompts for German are in tables 19 and 20.

Prompt template (generation)
Fornisci una breve definizione della parola "w_i" nella frase data in input. Genera solo la definizione. Input: "S_k"
Italian prompt
Fornisci una breve definizione della parola "sforzo" nella frase data in input. Genera solo la definizione. Input: "Che sforzo fate per valutare i risultati del vostro programme?"

Table 13: Prompt for the Italian generation benchmark.

Prompt template (multiple choice)
Data la parola "w_i" nella frase in input, scegli il significato corretto tra i seguenti: C_i. Genera solo il numero dell'opzione selezionata. Input: "S_k"
Italian prompt
Data la parola "valutare" nella frase in input, scegli il significato corretto tra i seguenti: 1) Esaminare o ascoltare (prove o un intero caso) per via giudiziaria. 2) Fare la stima commerciale di qlco. 3) Assegnare un valore a. 4) Ritenere dopo valutazione. 5) Apprezzare, tenere in grande stima. 6) Avere una certa opinione di qualcuno. Genera solo il numero dell'opzione selezionata. Input: "Che sforzo fate per valutare i risultati del vostro programme?"

Table 14: Prompt for the Italian multiple choice benchmark.

Prompt template (generation)
Proporciona una definición breve de la palabra "w_i" en la frase dada en entrada. Genera solo la definición. Input: "S_k"
Spanish prompt
Proporciona una definición breve de la palabra "reducido" en la frase dada en entrada. Genera solo la definición. Input: "¿ Mida su relación con el absentismo reducido, el volumen de negocios, los accidentes y las quejas, y con la mejora de la calidad y la producción?"

Table 15: Prompt for the Spanish generation benchmark.

Prompt template (multiple choice)
Dada la palabra "w_i" en la frase de entrada, elija el significado correcto entre los siguientes: C_i. Genera solo el número de la opción seleccionada. Input: "S_k"
Spanish prompt
Dada la palabra "esfuerzo" en la frase de entrada, elija el significado correcto entre los siguientes: 1) Actividad seria y consiente para hacer o lograr algo. 2) utilización de la fuerza y de otros medios por encima de lo normal con el fin de lograr un determinado objetivo 3) Ejercicio intenso o violento. 4) Enérgico intento de conseguir algo. 5) Intento que requiere un esfuerzo para conseguir un objetivo. Genera solo el número de la opción seleccionada. Input: "¿ Qué esfuerzo hace para evaluar los resultados de su programa?"

Table 16: Prompt for the Spanish multiple choice benchmark.

Prompt template (generation)
Donnez une brève définition du mot "w_i" dans la phrase d'entrée donnée. Ne donnez que la définition. Input: "S_k"
French prompt
Donnez une brève définition du mot "production" dans la phrase d'entrée donnée. Ne donnez que la définition. Input: "Mesurez -vous son rapport à la réduction de l'absentéisme, de chiffre d'affaires, des accidents et des griefs, ainsi qu'à l'amélioration de la qualité et de la production?"

Table 17: Prompt for the French generation benchmark.

Prompt template (multiple choice)
Étant donné le mot "w_i" dans la phrase saisie, choisissez la signification correcte parmi les suivantes: C_i. Ne donnez que le numéro de l'option sélectionnée. Input: "S_k"
French prompt
Étant donné le mot "essayez" dans la phrase saisie, choisissez la signification correcte parmi les suivantes: 1) Mettre à l'essai. 2) S'exercer à faire ou à effectuer quelque chose. 3) Tester l'apparence et la taille de (un vêtement) en le portant. Ne donnez que le numéro de l'option sélectionnée. Input: "Lorsque de améliorations sont recommandées dans les conditions de travail - comme l'éclairage, les salles de repos, les restaurants, la climatisation - essayez -vous de déterminer leur efficacité sur la productivité?"

Table 18: Prompt for the French multiple choice benchmark.

Prompt template (generation)
Geben Sie eine kurze Definition des Wortes "w_i" in dem gegebenen Satz an. Erzeugen Sie nur die Definition. Input: "S_k"
German prompt
Geben Sie eine kurze Definition des Wortes "Ziele" in dem gegebenen Satz an. Erzeugen Sie nur die Definition. Input: "Erreicht sie diese Ziele?"

Table 19: Prompt for the German generation benchmark.

Prompt template (multiple choice)
Wählen Sie für das Wort "w_i" im Eingabesatz die richtige Bedeutung aus den folgenden Angaben: C_i. Erzeugt nur die Nummer der ausgewählten Option. Input: "S_k"
German prompt
Wählen Sie für das Wort "Wahl" im Eingabesatz die richtige Bedeutung aus den folgenden Angaben: 1) Die Auswahl von etwas aus mehreren Möglichkeiten oder Alternativen. 2) Ein Stimmzettel, auch Wahlzettel, ist ursprünglich ein Zettel, auf dem der Wähler seine Wahl handschriftlich kundtun kann. 3) Weiler in Russland 4) Eine Wahl im Sinne der Politikwissenschaft ist ein Verfahren in Staaten, Gebietskörperschaften und Organisationen zur Bestellung einer repräsentativen Person oder mehrerer Personen als entscheidungs- oder herrschaftsausübendes Organ. Erzeugt nur die Nummer der ausgewählten Option. Input: "Stellen Sie bei Verhandlungen mit Ihrer Gewerkschaft sicher, dass die Mitarbeiter die Wahl zwischen neuen Leistungen und ihren Cents pro Stunde Lohnkosten haben."

Table 20: Prompt for the German multiple choice benchmark.