An Evaluation Framework for Long-tail Senses in Large Language Models and Word Sense Disambiguation

Junwei Zhang, Tianheng Wang, Tao Huang, Yupeng Zhang, Pengju Yan(✉), and Xiaolin Li(✉)

Center for AI and Intelligent Medicine, Hangzhou Institute of Medicine,
Chinese Academy of Sciences, Zhejiang Province, 310018, China.
zhangjunwei@him.cas.cn yanpengju@gmail.com xiaolinli@ieee.org

Abstract. In the field of Natural Language Processing (NLP), an excellent evaluation framework is crucial to the development of target tasks or fields. The emergence of Large Language Models (LLMs) requires some effective evaluation frameworks to protect their development. Traditional NLP tasks can be used to test the ability to understand and use language, but an evaluation framework that can test the accuracy and conciseness of language expression is still needed, that is, a dataset that can evaluate the ability to use long-tail senses. In addition, the Word Sense Disambiguation (WSD) task has benefited from some excellent evaluation frameworks and has been steadily developed. As models effectively identify high-frequency senses, the research focus of the WSD task has shifted to the identification of low-frequency senses, that is, long-tail senses. Based on the original evaluation framework of WSD, this paper constructs an evaluation framework that distinguishes high- and low-frequency senses, and the evaluation framework can be used to evaluate the vocabulary-level language understanding and expression ability of long-tail senses of LLMs, as well as the recognition ability of long-tail senses of WSD models.

Keywords: Evaluation Framework · Word Sense Disambiguation · Large Language Models.

1 Introduction

In the field of Machine Learning (ML), an evaluation framework refers to a collection of datasets and metrics used to evaluate model performance. An excellent evaluation framework can objectively and effectively evaluate the effectiveness and contribution of the proposed model, and promote fair competition among models [1, 41]. The rapid and efficient development of many subtasks in the field of Natural Language Processing (NLP) is due to the corresponding evaluation framework [11, 39, 40], such as the dataset WMT (Workshop on Statistical Machine Translation) and the evaluation metric BLEU (Bilingual Evaluation Understudy) score in machine translation tasks [12].

Large Language Models (LLMs) have risen rapidly in 2023 and have quickly become a research hotspot in the field of Artificial Intelligence [6]. It is foreseeable that LLMs will have an unprecedented and far-reaching impact in many fields such as natural language understanding and generation, text generation and creation, cross-language communication, intelligent assistants, etc., and will bring new opportunities and challenges. Currently, LLMs are evaluated on many traditional tasks in the field of NLP, such as natural language understanding, machine translation, dialogue systems, text generation, code generation, question and answer tasks, and language model ability testing, but are rarely evaluated at the vocabulary-level language understanding and expression ability of LLMs. **The evaluation framework at the lexical level is conducive to testing the understanding of the vocabulary itself, the richness of the words, and the language expression ability of LLMs, and is beneficial to promoting the development of LLMs.**

In addition, Word Sense Disambiguation (WSD) is the most basic research topic in the field of NLP [17, 37]. The WSD task aims to determine the most likely gloss in a list of word sense definitions for the target word based on given contextual information, and is a standard classification task. With the development and progress of machine learning and word sense recognition technology, the recognition accuracy of high-frequency word senses, that is, commonly used word senses, has reached expectations. Therefore, the current research focus turns to the most difficult low-frequency word senses (also known as long-tail word senses), that is, long-tail WSD [35, 38, 36]. However, **there is no evaluation framework suitable for evaluating long-tail WSD tasks in the community**.

For the above two motivations, based on the evaluation framework of WSD proposed by Raganato et al. [22], called the original evaluation framework, this paper proposes an evaluation framework that distinguishes high- and low-frequency word senses, in order to promote the development of long-tail word sense disambiguation tasks and LLMs. Because it is difficult to distinguish between high- and low-frequency word senses, that is, the identification criteria vary from person to person, we use machine and manual methods to implement it. The cooperation between man and machine can avoid the inevitable negligence or mistakes caused by workers in the tedious work process. First, we leverage ChatGPT 4.0 to identify high- and low-frequency word senses for target words in the given text, in which two different prompting projects are used to repeatedly identify the target words; Then, target words with inconsistent recognition results given by the machine method are manually screened by multiple people to provide a final division of high- and low-frequency word senses.

The contributions are summarized as follows:

- An evaluation framework of WSD suitable for evaluating long-tail word sense disambiguation tasks and LLMs is constructed using machine and manual methods. The framework has been published online, https://qnlp.github.io/.
- Based on the evaluation framework, experiments under three settings are implemented on mainstream LLMs to testing the understanding of the vo-

cabulary itself, the richness of the words, and the language expression ability of LLMs.
– Based on the evaluation framework, experiments are conducted on previous WSD models to verify their ability to identify long-tail word senses.

2 Related Work

2.1 Evaluation Frameworks for WSD

The evaluation framework for WSD tasks has also gone through several stages of development, from the initial one that was only suitable for machine learning models with a small number of learning parameters, to the one that was suitable for deep learning models with a large number of learning parameters, and from the initial one that was only for the English language, to the one that was for multiple languages.

In the machine learning period, SemCor constructed by Miller et al. [14], OMSTI constructed by Taghipour et al. [28], Senseval-2 constructed by Edmonds et al. [9], Senseval-3 constructed by Snyder et al. [26], SemEval-07 constructed by Pradhan et al. [21], SemEval-13 constructed by Navigli et al. [18] and SemEval-15 constructed by Moro et al. [16] are used in WSD tasks. In the deep learning period, Navigli et al. [22] integrated WSD datasets commonly used in the machine learning period and constructed them into a general evaluation framework for WSD tasks. This evaluation framework has been widely adopted by the WSD community in the following 4 to 5 years, playing an important role in the development of the field. In addition, Pasini et al. [20] released a cross-lingual evaluation framework for WSD tasks in 2021, which has sense-annotated development and test sets in 18 languages from six different linguistic families, as well as language-specific training sets. This evaluation framework not only expands the size of datasets and provides conditions for training and fine-tuning pre-trained language models or LLMs, but the cross-lingual characteristics promote the transition of WSD models to multi-language scenarios.

The evaluation framework for WSD integrated by Navigli et al. [22] laid the foundation for the early development of WSD tasks. However, as the accuracy of recognition of high-frequency word senses by WSD models continues to improve, this evaluation framework is no longer applicable. Our work is to further improve this evaluation framework and develop it into an evaluation framework that can test the accuracy of long-tail word senses. Specifically, we divide the word senses in this evaluation framework into high-frequency and low-frequency word senses.

2.2 Evaluation Frameworks for LLMs

LLMs are standard generative pre-trained language models [22], and many tasks in the field of NLP can be handled by generating text, such as natural language understanding, machine translation, dialogue systems, text generation, code generation, etc. Therefore, the evaluation framework for the above tasks can be used to evaluate the performance of LLMs. Commonly used evaluation frameworks include GLUE (General Language Understanding Evaluation) [32], SuperGLUE [31], and SQuAD (Stanford Question Answering Dataset) [23]. GLUE is a multi-task evaluation framework that includes multiple natural language understanding tasks, such as text classification, sentiment analysis, natural language inference, etc., and can be used to evaluate the general performance of models. SuperGLUE is an extension of GLUE, including some more complex and challenging tasks, such as reading comprehension, multiple-choice questions, etc. SQuAD is a common evaluation framework for question-answering tasks, which consists of questions about an article, and the model needs to select answers from the article. It evaluates the model by calculating how much the answer generated based on the model overlaps with the standard answer. In addition, some benchmark tests can be used to understand the creativity and reasoning capabilities of LLMs, such as LAMBADA [19]. LAMBADA tests the contextual understanding capabilities of LLMs, requiring the model to predict missing words in a given context.

It is important to emphasize that LLMs have not yet been tested on vocabulary-level tasks. LLMs are pre-trained on a large number of common datasets and have the ability to understand high-frequency word senses. Testing the performance of LLMs on low-frequency word senses, namely, long-tail word senses, can better evaluate their language understanding and expression capabilities.

3 The Evaluation Framework for Long-tail Word Senses

3.1 The Original Evaluation Framework

Navigli et al. [22] released the first evaluation framework[1] for WSD suitable for deep learning models in 2017, which standardized the WSD task and promoted the development of the WSD community. The evaluation framework contains two training sets, SemCor [14] and OMSTI [28], and five test sets, Senseval-2 [9], Senseval-3 [26], SemEval-07 [21], SemEval-13 [18] and SemEval-15 [16], of which SemEval-07 is often used as a development set. SemCor is a manually sense-annotated corpus, and its list of glosses comes from WordNet 1.6. OMSTI is a large corpus annotated with senses, and its list of glosses comes from WordNet 3.0. The lists of glosses for Senseval-2, Senseval-3, SemEval-07, SemEval-13 and SemEval-15 are from WordNet 1.7, WordNet 1.7.1, WordNet 2.1, WordNet 3.0 and WordNet 3.0 respectively.

The evaluation metric used by the WSD models under the evaluation framework uniformly adopts the F1 score at the percentage, and the form that appears in the paper is $F1\text{-}score(\%)$. In past WSD tasks, F1 score is also the most commonly used metric.

[1] http://lcl.uniroma1.it/wsdeval/

Table 1. Statistics of high- and low-frequency word senses of the datasets in the constructed evaluation framework are given, where *NOUN*, *VERB*, *ADJ* and *ADV* refer to nouns, verbs, adjectives and adverbs respectively, *Total* refers to the entire dataset, and dataset ALL is the union of datasets Senseval-2, Senseval-3, SemEval-07, SemEval-13 and SemEval-15.

	High-frequency Word Senses					**Low-frequency Word Senses**				
Datasets	*NOUN*	*VERB*	*ADJ*	*ADV*	*Total*	*NOUN*	*VERB*	*ADJ*	*ADV*	*Total*
Senseval-2	1,005	483	425	234	2,147	61	34	20	20	135
Senseval-3	840	533	328	12	1,713	60	55	22	18	137
SemEval-07	147	275	0	0	422	12	21	0	0	33
SemEval-13	1,541	0	0	0	1,541	103	0	0	0	103
SemEval-15	506	144	149	76	975	25	7	11	4	47
ALL	4,040	1,542	902	326	6,810	260	110	53	20	443
SemCor	75,592	83,133	29,323	17,788	205,836	11,410	5201	2,430	1,159	20,200

3.2 The Constructed Evaluation Framework

Based on the original evaluation framework, this paper constructs an evaluation framework that distinguishes high- and low-frequency word senses. We leverage machine and manual methods to classify high- and low-frequency word senses. First, we use the **machine method** to identify high- and low-frequency word senses of target words based on the given text; then, we use the **manual method** to identify target words that cannot be identified by the machine method.

- The machine method uses the current best generative language model ChatGPT 4.0 (i.e., GPT-4) to judge the word sense of the target word. We constructed two types of prompts to drive GPT-4 to identify target words, and obtained two results accordingly. When the two results are the same, the results given by GPT-4 are used as the division of high- and low-frequency word senses of the target word; when the two results are not the same, the manual method is used for further identification.

The first purpose of using two prompts to drive GPT-4 is to avoid possible inapplicability in special circumstances under a single prompt. The second purpose is not to rely too much on GPT-4, which allows GPT-4 to identify simple instances as much as possible, and leave the difficult ones to manual processing. The high-frequency and low-frequency divisions corresponding to each word sense definition in the second prompt are given by WordNet. We consider word senses whose "Frequency Count" is less than or equal to 2 to be low-frequency word senses[2].

- The manual method uses joint identification by multiple people to give the final result. We hired five college students to perform manual identification of target words that could not be identified by the machine method. Five

[2] http://wordnetweb.princeton.edu/perl/webwn/

college students each gave their own identification results, and the final result with a higher proportion are used as the division of high- and low-frequency word senses of the target word.

The first purpose of hiring five students is to give results that are as reliable as possible. The second purpose of choosing an odd number of students is to ensure that there is no tie.

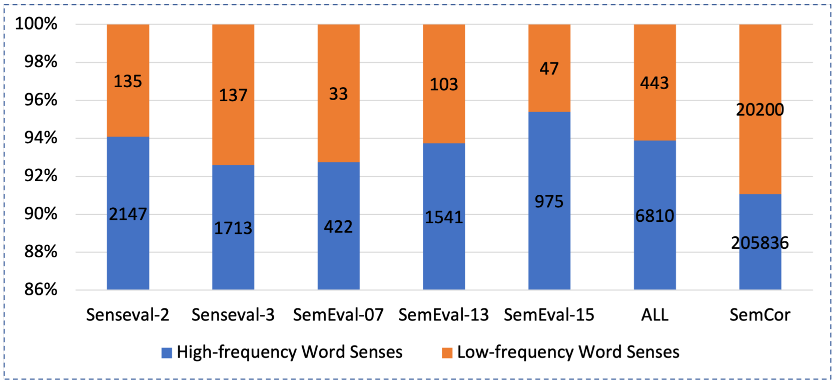

Fig. 1. The proportion of high- and low-frequency word senses in each dataset under the constructed evaluation framework.

The statistical information of the datasets in the constructed evaluation framework is shown in Tab. 1, in which the proportion of high- and low-frequency word senses in each dataset is shown in Fig. 1. As can be seen from Fig. 1, the proportion of low-frequency word senses (long-tail word senses) in all datasets is less than 10% and greater than 5%.

The evaluation metrics of the constructed evaluation framework also use the F1 score at the percentage used in the original evaluation framework. In addition, other content not presented in this paper remains consistent with the original evaluation framework.

4 Experiments and Analysis

In order to test the evaluation framework published in this paper, and also to evaluate the accuracy of the WSD models proposed by predecessors in identifying long-tail word senses, and the ability of LLMs to encode, understand and leverage long-tail word senses, the following two parts of experiments are conducted.

4.1 WSD under The Evaluation Framework

Experimental models: Representative WSD models in the past five years are used as experimental models, including LMMS [13], GlossBERT [10], BEM [5],

Table 2. Experimental results of WSD models under the constructed evaluation framework: *HF* and *LF* refer to the experimental results under high- and low-frequency word senses respectively.

	Senseval-2		Senseval-3		Semeval-07		Semeval-13		Semeval-15		ALL	
WSD Models	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*
LMMS	77.3	60.0	76.6	62.0	68.5	63.6	76.8	49.5	77.6	63.8	76.4	60.3
GlossBERT	79.5	50.4	77.3	58.4	73.0	60.6	78.1	57.3	79.8	68.1	78.2	56.9
BEM	80.9	55.6	78.5	64.2	75.8	57.6	80.9	62.1	82.4	68.1	80.2	60.3
ARES	79.0	63.0	78.5	59.9	71.8	60.6	79.0	51.5	83.4	78.7	79.0	60.9
KWSD	70.9	48.9	68.4	37.2	58.3	39.4	70.4	37.9	73.0	57.4	69.5	44.9
EWISER	79.8	64.4	79.5	65.0	71.6	63.6	80.1	61.2	79.8	68.1	79.3	63.2
Syntagrank	72.8	51.1	74.1	46.0	61.1	36.4	74.2	42.7	76.6	59.6	73.2	48.8
Generationary	79.0	59.3	75.0	58.4	69.2	63.6	79.9	53.4	78.3	63.8	77.4	59.1
ESC	82.4	71.1	78.9	64.2	76.8	66.7	83.3	67.0	83.4	78.7	81.5	68.8
ESR	81.2	63.0	79.9	60.6	76.5	57.6	81.4	58.3	82.6	78.7	80.7	63.4
SemEq-Base	82.6	64.4	80.6	63.5	75.4	63.6	83.1	54.4	82.3	78.7	81.6	63.9
SS-WSD	75.7	56.3	74.0	56.9	64.9	54.5	78.3	64.1	81.1	74.5	75.9	61.4

EWISER [4], KWSD [33], ARES [24], Syntagrank [25], Generationary [3], ESR [27], ESC [2], SemEq-Base [34] and SS-WSD [15]. In addition, the basis for selecting the above models is that the above WSD models all used the evaluation framework published by Navigli et al. [22] to implement experiments, that is, the original evaluation framework.

Results and Analysis: The experimental results of the models are obtained by counting the intermediate files published in the original papers. Based on the division of high- and low-frequency word senses in the constructed evaluation framework, the experimental results of high- and low-frequency word senses in the corresponding dataset are calculated respectively.

The experimental results are shown in Tab. 2. Experimental results show that the performance of all WSD models on long-tail word senses (that is, low-frequency word senses) is much lower than the performance on high-frequency word senses, and the difference is about 20 percentage points. And the recognition accuracy of long-tail word senses is in the range of 60 percentage points and below. This phenomenon shows that the current mainstream WSD models cannot effectively identify long-tail word senses, and also proves that it is necessary and meaningful to publish an evaluation framework that distinguishes high- and low-frequency word senses.

In addition, the recognition accuracy of high-frequency word senses is not high, falling below 80 percentage points. This phenomenon shows that even the high-frequency word sense disambiguation task is not a solved task as people say.

4.2 LLMs under The Evaluation Framework

Testing the encoding ability of LLMs: To test the ability of LLMs to encode or represent long-tail word senses, we construct WSD models based on LLMs to encode glosses. These experiments can test the learning ability of LLMs for low-resource tasks.

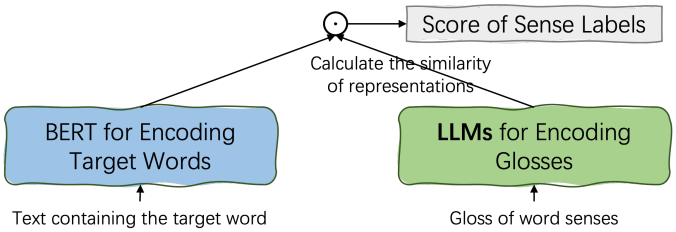
Score of Sense Labels
Calculate the similarity of representations
BERT for Encoding Target Words
LLMs for Encoding Glosses
Text containing the target word
Gloss of word senses

Fig. 2. WSD model based on LLMs to encode glosses to test the encoding or representation ability of LLMs.

Experimental models: The constructed WSD model is shown in Fig. 2. We use the pre-trained language model BERT [8] as the encoder to encode the target text containing the target word to obtain the representation of the target word, that is, the target word representation. The encoder that encodes the target word is called the target word encoder. We take the vector corresponding to the target word in the output of the encoder as the target word representation. If the target word contains multiple words, the corresponding output vectors are summed and averaged as the target word representation. We choose the mainstream LLMs, including LLaMA-1 [29], LLaMA-2 [30], LLaMA-3, Vicuna [7] and Falcon LLM [43], as the encoder to encode the glosses corresponding to the target word to obtain the representation of the word senses, that is, the word sense representation. The encoder that encodes the glosses is called the word sense encoder. We train the word sense encoder by adding a label $[CLS]$ to the end of the input text, and the vector corresponding to the label $[CLS]$ in the output of the encoder is regarded as the word sense representation. Then the similarity between the target word representation and each word sense representation is calculated, and the gloss corresponding to the highest similarity is the word sense of the target word.

It should be emphasized that the target word encoder obtains the word embedding of the target word, while the gloss encoder obtains the text embedding of the glosses. By using the above multiple LLMs as the word sense encoder, multiple experimental models are constructed.

Experimental settings: During the implementation process, the LLMs used in the above experimental models are fine-tuned. The training set used for fine-tuning is SemCor, the *learning rate* is 1e-5, the *context batch size* is 32, the *gloss batch size* is 128, and the *epoch* is 20. In addition, LLMs used to construct the experimental models respectively are the 7B and 13B versions of LLaMA-1, the 7B, Chat 7B, 13B and Chat 13B versions of LLaMA-2, the 8B and Chat 8B versions of LLaMA-3, the 7B and 13B versions of Vicuna, and the 7B version of Falcon LLM. The version used by the pre-trained language model BERT is *bert-base-uncased*. For other information about the models and hyperparameter settings that are not given in the paper, please see the code posted on the website where the evaluation framework is published.

Table 3. Experimental results of WSD models constructed by LLMs as the encoders under the constructed evaluation framework: *HF* and *LF* refer to the experimental results under high- and low-frequency word senses respectively.

	Senseval-2		Senseval-3		Semeval-07		Semeval-13		Semeval-15		ALL	
WSD Models	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*
LLaMA-1 7B	63.20	33.33	62.81	40.15	62.56	42.42	56.07	35.92	59.69	48.94	59.85	34.99
LLaMA-1 13B	58.92	30.37	62.52	37.23	57.35	27.27	52.63	27.18	59.18	48.49	58.43	32.96
LLaMA-2 7B	64.93	37.04	68.07	40.15	62.56	39.39	58.01	34.95	63.28	46.81	63.66	39.73
-chat 7B	63.95	36.30	69.59	43.07	62.09	48.48	58.53	37.86	64.10	48.94	63.95	41.76
LLaMA-2 13B	65.95	37.04	69.12	43.80	63.98	36.36	59.77	43.69	64.82	51.06	64.88	44.24
-chat 13B	66.28	38.52	69.47	45.26	63.51	39.39	59.57	41.75	64.21	51.06	64.96	44.02
LLaMA-3 8B	62.51	37.78	67.72	39.42	59.00	39.39	59.18	34.95	62.67	53.19	62.79	39.95
-chat 8B	62.65	37.04	67.43	37.23	57.35	39.39	58.14	33.98	61.95	53.19	62.29	39.28
Vicuna 7B	63.48	35.56	66.49	44.53	63.27	36.36	58.40	33.98	60.62	46.81	62.54	40.41
Vicuna 13B	64.93	37.04	69.59	43.07	63.51	36.36	58.79	37.86	65.23	48.94	64.48	42.44
Falcon 7B	62.51	40.00	66.55	38.69	60.19	33.33	57.75	30.10	63.79	53.19	62.45	38.15

Results and Analysis: The experimental results of the WSD models are shown in Tab. 3. The analysis of high- and low-frequency word sense recognition results is as follows:

From the overall performance, the recognition accuracy of high-frequency word senses remains between 50 and 70 percentage points, which is consistent with the results of models not based on LLMs. This phenomenon shows that LLMs have certain coding capabilities and can be used for text representation learning tasks. At the same time, it also shows that LLMs have certain encoding and understanding capabilities for high-frequency word senses. The shortcoming is that LLMs use a large amount of pre-training data but still do not improve the final recognition accuracy, indicating that a large amount of pre-training data has no substantial significance in improving the recognition of high-frequency word senses.

From the overall performance, the recognition accuracy of low-frequency word senses remains between 30 and 50 percentage points, which is far lower than the results that can be given by models not based on LLMs. This phenomenon shows that the ability of LLMs to encode or represent long-tail word senses is weak, and it further illustrates that there are still deficiencies in the language understanding ability of LLMs. It is understandable that the training data of

LLMs is some commonly used non-professional training corpus, in which the examples of long-tail word senses are lacking or account for a small proportion, resulting in the trained LLMs not having effective encoding and representation capabilities under low resource conditions.

Testing the understanding ability of LLMs: To test the ability of LLMs to understand long-tail word senses, we construct WSD models based on LLMs to generate glosses of target words. These experiments can test the language understanding ability of LLMs.

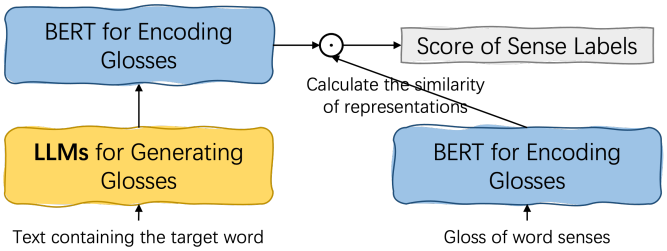
BERT for Encoding Glosses
Score of Sense Labels
Calculate the similarity of representations
LLMs for Generating Glosses
BERT for Encoding Glosses
Text containing the target word
Gloss of word senses

Fig. 3. WSD model based on LLMs to generate glosses to test the language understanding ability of LLMs.

Experimental models: The constructed WSD model is shown in Fig. 3. We choose the mainstream LLMs, including LLaMA-1 [29], LLaMA-2 [30], LLaMA-3, Vicuna [7], Gemma [42] and Falcon LLM [43], as the gloss generator of the target word to obtain the gloss based on the input text containing the target word, then use the pre-trained language model BERT [8] as the gloss encoder to separately encode the generated gloss and the known glosses to obtain the corresponding word sense representations, and finally calculate the similarity between the generated word sense representation and each known word sense representation, among which the gloss corresponding to the highest similarity is the word sense of the target word. The vector corresponding to the label $[CLS]$ output by the pre-trained language model BERT is used as a word sense representation.

It should be emphasized that the gloss encoder obtains the text embedding of the glosses. During the implementation process, the two gloss encoder do not share the same pre-trained language model BERT, but use their own pre-trained language model BERT respectively. By using the above multiple LLMs as the gloss generator, multiple experimental models are constructed.

Experimental settings: During the implementation process, the LLMs used in the above experimental models are fine-tuned. The training set used for fine-tuning is SemCor, the *learning rate* is 2e-6, the *batch size* is 4, and the *epoch* is 20. The pre-trained language model BERT also participates in fine-tuning, with the *learning rate* of 1e-5, the *context batch size* of 4, the *gloss batch* *size* of 256 and the *epoch* of 20. In addition, LLMs used to construct the experimental models respectively are the 7B and 13B versions of LLaMA-1, the 7B, Caht 7B, 13B and Chat 13B versions of LLaMA-2, the 8B and Caht 8B versions of LLaMA-3,the 7B and 13B versions of Vicuna, the 2B and 7B versions of Gemma and the 7B version of Falcon LLM. The version used by the pre-trained language model BERT is *bert-base-uncased*. For other information about the models and hyperparameter settings that are not given in the paper, please see the code posted on the website where the evaluation framework is published.

Table 4. Experimental results of WSD models based on LLMs to generate glosses under the constructed evaluation framework: *HF* and *LF* refer to the experimental results under high- and low-frequency word senses respectively.

	Senseval-2		Senseval-3		Semeval-07		Semeval-13		Semeval-15		ALL	
LLMs	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*
Gemma 2B	51.66	31.85	46.88	28.83	39.81	37.88	49.06	33.01	56.21	41.49	48.18	33.41
Gemma 7B	62.90	46.67	59.19	36.50	54.98	43.94	55.55	38.83	62.82	39.36	59.79	41.20
LLaMA-2 7B	68.82	42.96	68.16	39.42	63.74	45.45	58.92	48.54	71.64	62.77	66.52	43.00
-chat 7B	71.94	54.81	72.12	46.72	59.60	39.39	61.52	45.63	70.72	61.70	68.81	48.87
LLaMA-2 13B	68.56	48.52	67.25	45.99	61.61	43.94	61.71	46.60	69.38	62.77	66.40	46.50
-chat 13B	72.17	52.96	70.72	47.45	61.73	37.88	62.49	47.57	70.46	61.70	68.98	49.55
LLaMA-3 8B	59.25	33.33	59.05	40.15	49.88	36.36	50.68	33.98	61.08	57.45	56.75	37.70
-chat 8B	69.03	47.78	68.21	45.26	52.49	45.45	50.68	33.98	66.41	57.45	54.05	33.18
Vicuna 7B	59.43	42.22	58.61	35.04	54.50	40.91	54.32	33.01	62.31	37.23	57.63	37.13
Vicuna 13B	52.07	37.41	50.73	31.02	42.77	28.79	49.06	36.89	55.85	40.43	51.23	36.12
Falcon 7B	59.85	39.26	55.81	31.02	52.96	33.33	55.55	33.01	64.97	53.19	58.73	37.25

Results and Analysis: The experimental results of the WSD models are shown in Tab. 4. The analysis of high- and low-frequency word sense recognition results is as follows:

From the overall performance, the recognition accuracy of high-frequency word senses remains between 50 and 70 percentage points, which is lower than the results of models not based on LLMs, and also lower than the results of models with LLMs as encoders. This phenomenon shows that LLMs have a certain ability to understand high-frequency word senses, but it cannot effectively provide word sense description information, that is, word sense definition, which further illustrates that the language expression ability of LLMs is insufficient. We provide further experimental analysis on this issue in the next section. In fact, giving a definition of word sense is a difficult task. It not only requires LLMs to have an accurate understanding of word sense, but also requires LLMs to have the ability to accurately summarize.

From the overall performance, the recognition accuracy of low-frequency word senses remains between 30 and 50 percentage points, which is much lower than the results of models not based on LLMs, and is consistent with the results of models with LLMs as encoders. This phenomenon shows that the ability of LLMs to understand long-tail word senses is weak, and it also reflects that the ability of LLMs to define and summarize long-tail word senses is weak. The reason why LLMs have a weak ability to understand long-tail word senses is that the training data for training LLMs contains fewer examples of long-tail word senses, and at the same time, the small proportion of low-resource training data is ignored during the model training process. The weak ability of LLMs to summarize the definition of long-tail word senses is also due to the fact that the training data of LLMs contains fewer examples of long-tail word senses, so they cannot summarize the definitions of long-tail word senses.

Testing the expression ability of LLMs: To test the ability of LLMs to leverage long-tail word senses, we construct WSD models based on LLMs to generate example sentences with target words. These experiments can test the language expression ability of LLMs.

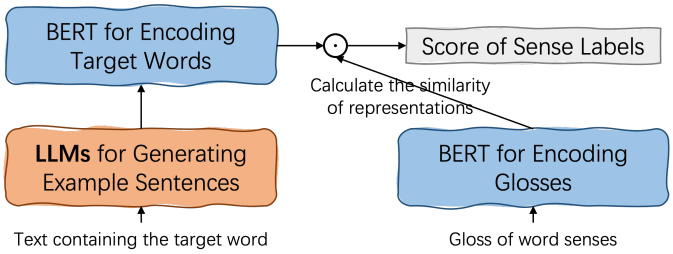
BERT for Encoding Target Words
Score of Sense Labels
Calculate the similarity of representations
LLMs for Generating Example Sentences
BERT for Encoding Glosses
Text containing the target word
Gloss of word senses

Fig. 4. WSD model based on LLMs to generate example sentences to test the language expression ability of LLMs.

Experimental models: The constructed WSD model is shown in Fig. 4. We choose the mainstream LLMs, including Gemma [42], LLaMA-2 [30], LLaMA-3, Vicuna [7] and Falcon LLM [43], as the example sentence generator of the target word to obtain the example sentence based on the input text containing the target word, then use the pre-trained language model BERT [8] as the encoder to encode the generated the example sentence and the glosses to obtain the corresponding target word representation and gloss representations, and finally calculate the similarity between the target word representation and each word sense representation, among which the gloss corresponding to the highest similarity is the word sense of the target word. The vector corresponding to the label $[CLS]$ output by the pre-trained language model BERT is used as a word sense representation. The vector corresponding to the target word output by the pre-trained language model BERT is used as the target word representation. If the target word contains multiple words, the corresponding output vectors are summed and averaged as the target word representation.

It should be emphasized that the target word encoder obtains the word embedding of the target word, while the word sense encoder obtains the text embedding of the glosses. During the implementation process, the two encoder do not share the same pre-trained language model BERT, but use their own pre-trained language model BERT respectively. By using the above multiple LLMs as the example sentence generator, multiple experimental models are constructed.

Experimental settings: During the implementation process, the LLMs used in the above experimental models are fine-tuned. The training set used for fine-tuning is SemCor, the *learning rate* is 2*e*-6, the *batch size* is 4, and the *epoch* is 20. The pre-trained language model BERT also participates in fine-tuning, with the *learning rate* of 1e-5, the *context batch size* of 4, the *gloss batch size* of 256 and the *epoch* of 20. In addition, LLMs used to construct the experimental models respectively are the 2B and 7B versions of Gemma, the 7B, Chat 7B, 13B and Chat 13B versions of LLaMA-2, the 8B and Chat 8B versions of LLaMA-3, the 7B and 13B versions of Vicuna, and the 7B version of Falcon LLM. The version used by the pre-trained language model BERT is *bert-base-uncased*. For other information about the models and hyperparameter settings that are not given in the paper, please see the code posted on the website where the evaluation framework is published.

Table 5. Experimental results of WSD models based on LLMs to generate example sentences under the constructed evaluation framework: *HF* and *LF* refer to the experimental results under high- and low-frequency word senses respectively.

	Senseval-2		Senseval-3		Semeval-07		Semeval-13		Semeval-15		ALL	
LLMs	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*	*HF*	*LF*
Gemma 2B	43.27	23.53	56.66	34.55	52.36	38.10	54.77	32.04	55.33	57.14	51.30	34.55
Gemma 7B	55.28	26.47	60.41	40.00	58.55	38.10	60.09	40.78	58.20	57.14	57.72	40.91
LLaMA-2 7B	53.21	29.41	57.22	32.73	64.36	33.33	55.81	30.10	52.05	42.86	54.02	30.91
-chat 7B	51.97	29.41	61.73	36.36	60.73	42.86	70.08	43.69	59.43	57.14	57.91	38.18
LLaMA-2 13B	54.45	35.29	57.60	27.27	63.27	47.62	56.33	29.13	56.15	57.14	55.64	27.27
-chat 13B	56.52	32.35	64.35	43.64	61.82	38.10	59.25	36.89	63.93	57.14	61.02	43.64
LLaMA-3 8B	72.75	50.37	67.13	42.34	66.59	42.42	66.39	35.92	67.18	48.94	67.67	44.02
-chat 8B	72.15	53.33	67.89	37.23	63.98	42.42	66.90	34.95	69.03	46.81	68.16	41.53
Vicuna 7B	54.04	32.35	62.85	36.36	62.55	42.86	59.25	37.86	60.25	57.14	59.27	40.91
Vicuna 13B	55.07	32.35	63.79	45.45	63.27	38.10	60.42	36.89	61.48	57.14	60.38	42.73
Falcon 7B	67.44	46.67	66.14	45.26	57.35	39.39	65.35	37.86	67.90	46.81	66.09	42.89

Results and Analysis: The experimental results of the WSD models are shown in Tab. 5. The analysis of high- and low-frequency word sense recognition results is as follows:

This part of the experiment can test the ability of LLMs to understand and leverage high-frequency word senses. However, from the overall performance, the LLMs are still unable to leverage high-frequency word senses well in complex scenarios. This phenomenon shows that in complex scenarios, LLMs still cannot correctly understand and leverage common word senses at the vocabulary level. It reveals that LLMs only generate content based on language inertia and do not have a deep understanding of the training data. This phenomenon also shows that in future research work, it is necessary to enable LLMs to better learn the inherent meaning of language from the lexical level, so that LLMs have excellent language understanding and expression capabilities.

This part of the experiment can test the ability of LLMs to understand and leverage long-tail word senses, but from the overall performance, the LLMs have not yet reached expectations. We believe that giving LLMs the ability to understand and leverage long-tail word senses requires research starting from the core structure of the model. The current LLMs based on pre-training methods will inevitably lead to the neglect of long-tail word senses, because long-tail word senses are extremely scarce in the training data, and it is impossible to make LLMs pay attention to low-resource samples during the training phase.

5 Conclusions

To promote and facilitate the development of long-tail Word Sense Disambiguation (WSD) and Large Language Models (LLMs), this paper releases an evaluation framework that distinguishes high- and low-frequency word senses. For WSD, the framework can distinguish the recognition accuracy of the WSD models in identifying high- and low-frequency word senses, and then design the models in a targeted manner. For LLMs, the framework can detect the word richness, language understanding and expression capabilities of the language models at the vocabulary level, and then design solutions with clear goals. Based on the framework, this paper tests the accuracy of the WSD models proposed by previous researchers for long-tail sense recognition. The results show that the long-tail WSD task still has a long way to go. And this paper tests the ability of LLMs to encode, understand, and leverage long-tail senses under the framework. The results show that LLMs cannot effectively deal with long-tail word senses.

Acknowledgements Our work is supported by the National Key Research and Development Program of China (No. 2022YFC3600902) and the National Science Technology Major Project (No. 2018AAA0101904, 2023ZD0509706).

References

1. Alpaydin, E.: Machine learning. MIT press (2021)
2. Barba, E., Pasini, T., Navigli, R.: ESC: redesigning WSD with extractive sense comprehension. In: NAACL. pp. 4661–4672 (2021)
3. Bevilacqua, M., Maru, M., Navigli, R.: Generationary or “how we went beyond word sense inventories and learned to gloss". In: EMNLP. pp. 7207–7221 (2020)
4. Bevilacqua, M., Navigli, R.: Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In: ACL. pp. 2854–2864 (2020)

5. Blevins, T., Zettlemoyer, L.: Moving down the long tail of word sense disambiguation with gloss-informed biencoders. CoRR **abs/2005.02590** (2020)
6. Cerf, V.G.: Large language models. Commun. ACM **66**(8), 7 (2023)
7. Chiang, W., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (2023)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Edmonds, P., Cotton, S.: SENSEVAL-2: overview. In: Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL@ACL 2001. pp. 1–5 (2001)
10. Huang, L., Sun, C., Qiu, X., Huang, X.: Glossbert: BERT for word sense disambiguation with gloss knowledge. In: EMNLP. pp. 3507–3512 (2019)
11. Jones, K.S., Galliers, J.R.: Evaluating natural language processing systems: An analysis and review (1995)
12. Lopez, A.: Statistical machine translation. ACM Computing Surveys (CSUR) **40**(3), 1–49 (2008)
13. Loureiro, D., Jorge, A.: Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In: ACL. pp. 5682–5691 (2019)
14. Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: Human Language Technology, Proceedings of a Workshop held at Plainsboro (1994)
15. Mizuki, S., Okazaki, N.: Semantic specialization for knowledge-based word sense disambiguation. In: EACL. pp. 3449–3462 (2023)
16. Moro, A., Navigli, R.: Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). pp. 288–297 (2015)
17. Navigli, R.: Word sense disambiguation: A survey. ACM computing surveys (CSUR) **41**(2), 1–69 (2009)
18. Navigli, R., Jurgens, D., Vannella, D.: Semeval-2013 task 12: Multilingual word sense disambiguation. In: Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013. pp. 222–231 (2013)
19. Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q.N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., Fernández, R.: The lambada dataset: Word prediction requiring a broad discourse context. arXiv preprint arXiv:1606.06031 (2016)
20. Pasini, T., Raganato, A., Navigli, R.: XL-WSD: an extra-large and cross-lingual evaluation framework for word sense disambiguation. In: AAAI. pp. 13648–13656 (2021)
21. Pradhan, S., Loper, E., Dligach, D., Palmer, M.: Semeval-2007 task-17: English lexical sample, SRL and all words. In: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007. pp. 87–92 (2007)
22. Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: EACL. pp. 99–110 (2017)
23. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
24. Scarlini, B., Pasini, T., Navigli, R.: With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In: EMNLP. pp. 3528–3539 (2020)

25. Scozzafava, F., Maru, M., Brignone, F., Torrisi, G., Navigli, R.: Personalized pagerank with syntagmatic information for multilingual word sense disambiguation. In: ACL. pp. 37–46 (2020)
26. Snyder, B., Palmer, M.: The english all-words task. In: Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, SENSEVAL@ACL 2004 (2004)
27. Song, Y., Ong, X.C., Ng, H.T., Lin, Q.: Improved word sense disambiguation with enhanced sense representations. In: EMNLP. pp. 4311–4320 (2021)
28. Taghipour, K., Ng, H.T.: One million sense-tagged instances for word sense disambiguation and induction. In: ACL. pp. 338–344 (2015)
29. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023)
30. Touvron, H., Martin, L., Stone, K.: Llama 2: Open foundation and fine-tuned chat models. CoRR **abs/2307.09288** (2023)
31. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems **32** (2019)
32. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multitask benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
33. Wang, Y., Wang, M., Fujita, H.: Word sense disambiguation: A comprehensive knowledge exploitation framework. Knowl. Based Syst. **190**, 105030 (2020)
34. Yao, W., Pan, X., Jin, L., Chen, J., Yu, D., Yu, D.: Connect-the-dots: Bridging semantics between words and definitions via aligning word sense inventories. In: EMNLP. pp. 7741–7751 (2021)
35. Zhang, J., He, R., Guo, F.: Bi-matching mechanism to combat long-tail senses of word sense disambiguation. In: ECML-PKDD. Lecture Notes in Computer Science, vol. 13714, pp. 621–637 (2022)
36. Zhang, J., He, R., Guo, F.: Quantum-inspired representation for long-tail senses of word sense disambiguation. In: AAAI. pp. 13949–13957 (2023)
37. Zhang, J., He, R., Guo, F., Liu, C.: Quantum interference model for semantic biases of glosses in word sense disambiguation. In: AAAI. pp. 19551–19559 (2024)
38. Zhang, J., He, R., Guo, F., Ma, J., Xiao, M.: Disentangled representation for long-tail senses of word sense disambiguation. In: CIKM. pp. 2569–2579 (2022)
39. Zhang, J., Hou, Y., Li, Z., Zhang, L., Chen, X.: Strong statistical correlation revealed by quantum entanglement for supervised learning. In: ECAI. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 1650–1657 (2020)
40. Zhang, J., Li, Z., Wang, J., Wang, Y., Hu, S., Xiao, J., Li, Z.: Quantum entanglement inspired correlation learning for classification. In: PAKDD. vol. 13281, pp. 58–70 (2022)
41. Zhang, J., Wang, T., Wang, C., Bai, Y., Zhang, Y., Li, Y.: Emotional polarity attention mechanism for text sentiment analysis. In: DASFAA. vol. 14854, pp. 3–18 (2024)
42. Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R.C., Van Rossum, T., McDonald, C., Hall, A., Wan, X., Lim, R., et al.: Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics **28**(17), 2272–2273 (2012)
43. ZXhang, Y.X., Haxo, Y.M., Mat, Y.X.: Falcon llm: A new frontier in natural language processing. AC Investment Research Journal **220**(44) (2023)