# An Exploration-Analysis-Disambiguation Reasoning Framework for Word Sense Disambiguation with Low-Parameter LLMs

**Deshan Sumanathilaka, Nicholas Micallef, Julian Hough**
Department of Computer Science, Swansea University
Fabian Way, Crymlyn Burrows, Swansea, Wales, UK
{t.g.d.sumanathilaka, nicholas.micallef, julian.hough}@swansea.ac.uk

**Abstract**

Word Sense Disambiguation (WSD) remains a key challenge in Natural Language Processing (NLP), especially when dealing with rare or domain-specific senses that are often misinterpreted. While modern high-parameter Large Language Models (LLMs) such as GPT-4-Turbo have shown state-of-the-art WSD performance, their computational and energy demands limit scalability. This study investigates whether low-parameter LLMs (<4B parameters) can achieve comparable results through fine-tuning strategies that emphasize reasoning-driven sense identification. Using the FEWS dataset augmented with semi-automated, rationale-rich annotations, we fine-tune eight small-scale open-source LLMs (e.g. Gemma and Qwen). Our results reveal that Chain-of-Thought (CoT)-based reasoning combined with neighbour-word analysis achieves performance comparable to GPT-4-Turbo in zero-shot settings. Importantly, Gemma-3-4B and Qwen-3-4B models consistently outperform all medium-parameter baselines and state-of-the-art models on FEWS, with robust generalization to unseen senses. Furthermore, evaluation on the unseen "Fool Me If You Can" dataset confirms strong cross-domain adaptability without task-specific fine-tuning. This work demonstrates that with carefully crafted reasoning-centric fine-tuning, low-parameter LLMs can deliver accurate WSD while substantially reducing computational and energy demands.

**Keywords:** Word Sense Disambiguation, Low-parameter LLMs, Reasoning-driven Fine-tuning

## 1. Introduction

With recent advances in large language models (LLMs), Natural Language Processing (NLP) tasks such as machine translation, information retrieval, question answering, sentiment analysis, and summarization have achieved near-human performance in modern applications (Chang et al., 2024; Sumanathilaka et al., 2025c). However, to perform such linguistic tasks effectively, word sense disambiguation (WSD) plays a major role, as ambiguity can lead to incorrect information and the generation of misinformation (Singh et al., 2024). For example, during the COVID-19 pandemic, the term "*positive*" could indicate a confirmed infection or express optimism in a general context which are two drastically different meanings that require contextual understanding to avoid incorrect inferences. Understanding the meaning of a word in a sentence or phrase and disambiguating the correct sense is crucial for achieving accurate results, particularly when tackling the lexical ambiguity of an ambiguous word.

Working with words that have multiple meanings (polysemy) is a major challenge for current NLP systems, as highly polysemous words are difficult to disambiguate due to the nature of usage. For example, according to BabelNet[1], the word '*bank*' has 49 meanings, including 40 as a nouns and 9 as a verb. However, in common usage, the senses "financial institution" (money bank) and "sloping land beside a body of water" (river bank) dominate. In contrast, senses including "a flight maneuver" (highly domain-specific to aviation) or "a slope in the turn of a road or track" (used mainly in engineering, road design) are rare in everyday usage. In domain-specific situations, models often struggle to disambiguate less common senses effectively (Awotunde, 2025).

Recent studies on WSD using large language models (LLMs) have revealed promising findings, including significant improvements in disambiguating rare senses and leveraging broader contextual information, paving the way for enhanced language understanding and more accurate disambiguation. For instance, Sumanathilaka et al. (2024b) evaluated multiple large-parameter LLMs for disambiguation using few-shot and knowledge-base-combined methods, highlighting the strengths of LLMs for WSD. However, other studies also reveal important limitations. Basile et al. (2025) demonstrated that while LLMs perform well in zero-shot scenarios, they cannot surpass current state-of-the-art methods without fine-tuning, suggesting that existing approaches do not fully exploit LLM capabilities or generalize across diverse WSD settings. These gaps highlight the need for research that addresses scalability, efficiency, and more effective use of LLMs for robust and generalizable disambiguation. In particular, incorporating proper inner reasoning to guide decision making is crucial, motivating the present study on reasoning-driven WSD with low-

[1] https://babelnet.org/



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10390–10404
11–16 May 2026. 

parameter LLMs. Inspired by these observations, we design and evaluate a smaller, energy-efficient LLMs for WSD with three primary design strategies:

1. Given a word in a sentence, the LLM must generate the correct definition.
2. Given a word in a sentence and possible senses, the model reasons using the neighbouring context closely related to the disambiguation process to identify the correct sense.
3. Given a word in a sentence and possible senses, the model reasons why a particular sense is correct and why other senses are not, to determine the correct meaning.

Following these design principles, we fine-tune energy-efficient, low-parameter open-source models with fewer than 4B parameters (Bai et al., 2024). Our study reveals that a chain-of-thought (CoT)-based reasoning process combined with neighbouring context analysis performs comparably to large models like GPT-4-Turbo, achieving high performance on unseen data. Furthermore, our advanced reasoning approach requires only 10% of the training data compared to the second approach, yet performed similarly, demonstrating its effectiveness. The main contributions of this work are:

- Building three reasoning datasets using a semi-automated approach and releasing them as open-source resources for future research.[2] [3]
- Proposing and implementing lightweight adapters following a novel EAD (Exploration, Analysis and Disambiguation) framework that can be efficiently used for WSD with small-scale models.
- Evaluating eight open-source small-parameter models with the proposed approaches and achieving superior performance compared to all medium-parameter models.

The rest of the paper is organized as follows: background and related work, which discusses the WSD evolution and related studies; methodology, which outlines the approach used to conduct this study; results and observations, followed by conclusions and future directions.

## 2. Background and Related Work

### 2.1. Evolution of WSD

WSD has a long history, evolving from rule-based methods to deep learning and LLM approaches. Early rule-based systems relied on hand-crafted grammatical and syntactic rules to exploit the local context of ambiguous words (Bowerman, 1978), but these methods were labor-intensive, domain-specific, and difficult to scale to complex languages (Palmer et al., 2006). Knowledge-based (KB) approaches emerged with resources such as WordNet, dictionaries, and thesauri incorporating algorithms such as Lesk to measure the sense overlap between gloss definitions and the surrounding context (Lesk, 1986). Advanced KB solutions such as random walks over extended lexical graphs including Extended WordNet (Agirre et al., 2014) and graph-based methods that combined word embeddings with contextual information were explored (Duarte et al., 2021). Exploiting synset definitions, hypernymy relations, and contextual features improved the accuracy (Kolte and Bhirud, 2009), with frameworks like the Synset Relation-Enhanced Framework (SREF) achieving state-of-the-art KB WSD (Wang and Wang, 2020). Filtering unnecessary semantic information (Kwon et al., 2021) and incorporating dependency parsing to extract more precise contextual knowledge (Meng, 2022) were also prominent KB innovations.

With annotated corpora, supervised Machine Learning became the most common research approach (Le and Shimazu, 2004; Al-Bayaty and Joshi, 2016; Gosal, 2015). These methods exploited features, namely target word morphology, Part of Speech (POS) tags, syntactic dependencies, and collocations. To achieve efficient WSD, supervised neural architectures incorporated richer lexical and gloss-based information to the training pipeline. Gloss-augmented neural networks jointly encoded glosses and context were examined (Luo et al., 2018). Other approaches to supervised WSD explored multiple-sense identification (Orlando et al., 2021), stacked BiLSTMs with attention (Laatar et al., 2023), and context-dependent modeling (Koppula et al., 2021).

There was a major shift in WSD research with the arrival of Transformer architectures (Vaswani, 2017), which enabled deep, attention-based contextual modeling. Transformer-based models such as BERT and GPT achieved state-of-the-art performance on WSD tasks (Huang et al., 2020), demonstrating strong generalization to new domains due to extensive pre-training on diverse corpora. For instance, SenseBERT augments BERT pre-training by requiring the model to predict masked words and their WordNet supersenses (Levine et al., 2019), while GlossBERT constructs context–gloss pairs and fine-tunes BERT (Huang et al., 2020). Moreover, (Barba et al., 2021b,a; Blevins and Zettlemoyer, 2020) reshaped the field with supervised approaches that enhance WSD performance.

Current research focuses on zero-shot and few-shot WSD with LLMs, applying in-context learning to leverage linguistic knowledge without extensive retraining (Basile et al., 2025; Yae et al.,

[2] Reasoning dataset: https://huggingface.co/datasets/deshanksuman/Reasoning_WSD_dataset

[3] Reasoning dataset for Verb: https://huggingface.co/datasets/deshanksuman/Advance_Reasoning_for_Verb_WSD

2025). These approaches combine large-scale pre-training, dynamic context modeling, and KB augmentation, representing the most powerful and versatile WSD techniques to date and thus forming the basis of this study's methods.

### 2.2. Large Language Models for WSD

Despite their remarkable performance across numerous NLP tasks, LLMs still face several challenges in WSD. While they excel at handling common vocabulary, recent research indicates that LLMs often misinterpret rare or domain-specific ambiguous terms, especially in cross-lingual scenarios (Cahyawijaya et al., 2024; Yae et al., 2024; Basile et al., 2025; Ortega-Martín et al., 2023; Meconi et al., 2025). For example, Cahyawijaya et al. (2024) highlighted persistent errors between languages, with a bias toward higher-resource languages. Furthermore, Yae et al. (2024) observed that model size strongly influences WSD accuracy, yet larger models demand significantly more computational resources, raising efficiency concerns. Although more parameter-efficient language models generally lack the rich contextual understanding of their larger counterparts, Basile et al. (2025) demonstrates that fine-tuning them for downstream WSD tasks substantially improves accuracy over base configurations. This approach offers the potential for energy-efficient, domain-specialized disambiguation.

Recent work has explored various strategies for leveraging LLMs in WSD. Sainz et al. (2023) reframed WSD as a textual entailment task, prompting LLMs to evaluate domain label suitability for a sentence containing an ambiguous word. This zero-shot method not only outperformed random guessing but also, in certain cases, matched or exceeded supervised WSD systems (Ortega-Martín et al., 2023). Similarly, Sumanathilaka et al. (2024a) investigated prompt engineering with in-context learning in GPT-3.5-Turbo and GPT-4, while subsequent work benchmarked the WSD performance of multiple LLMs, identifying DeepSeek R1 and GPT-4-mini as effective (Sumanathilaka et al., 2024b).

Few studies have examined the affect of functional variables in LLMs. Mainly, temperature tuning was shown to influence disambiguation accuracy (Sumanathilaka et al., 2025a; Li et al., 2025b). Model architecture is crucial, with Qorib et al. (2024) reporting that encoder-only models can outperform decoder-only designs for this task. Meanwhile, multilingual and translation-based WSD strategies continue to be explored (Kang et al., 2023; David et al., 2024; Ren et al., 2024; Abdel-Salam, 2024; Laba et al., 2023), reflecting the ongoing relevance even with LLMs, as its role has shifted from a standard pipeline component to a targeted research focus.

Table 1: Stats for fine-tuning Variants. R: Reason

| Dataset | Size (K) | Input | | Output | |
|---|---|---|---|---|---|
| | | Max | Avg | Max | Avg |
| Direct sense | 101 | 511 | 44.7 | 251 | 13.9 |
| COT R | 101 | 1915 | 226.1 | 1921 | 265.9 |
| Advanced R | 10 | 1865 | 212.6 | 2477 | 672.0 |
| Verb R | 4.5 | 1915 | 222.6 | 2607 | 764.4 |

## 3. Methodology

With the aim of achieving a reduced use of computational, memory, energy, and financial resources, we present the methodology employed in our work using low-parameter models (<4B parameters) for an efficient reasoning driven WSD task.

### 3.1. Dataset and Augmentation Process

This work employs the FEWS dataset as the basis for the experiments, which includes the sense tag list, training data, and test data (Blevins et al., 2021). FEWS was chosen because it contains less frequently used ambiguous words compared to the Semcor and Unified Evaluation Framework datasets (Raganato et al., 2017). Additionally, the distribution of training and test data, designed for evaluation in both few-shot and zero-shot settings, makes it well-suited for a fair and meaningful comparison (Goworek et al., 2025). The training set contains ≈101K samples, and each test set has 5,000 records for evaluation.

We augment the FEWS training data to support the fine-tuning process (Blevins et al., 2021). Our fine-tuning method is designed to elicit chain-of-thought (CoT) reasoning together with neighbour-word analysis to select the closest semantic candidate to determine word sense (see subsection 3.2.2). For advanced reasoning, we employ the *Virtuoso-Large*, an open source model from Arcee.ai[4] to generate rationales for correct sense assignments and for rejecting competing senses. The model was selected based on its strong empirical performance reported in Sumanathilaka et al. (2024b), where it ranked among the best-performing models, and due to the availability of open weights, ensuring transparency and reproducibility. The model was instructed to produce a structured rationale based on three factors: contextual analysis, justification to the correct sense, and reasoning for elimination of incorrect sense. The process used a human-in-the-loop approach to ensure accurate data augmentation. We conducted additional experiments to evaluate computational techniques for verb disambiguation. To enhance

[4] https://huggingface.co/arcee-ai/Virtuoso-Large

the dataset, we incorporated syntactic evidence into the reasoning chain in addition to the semantic evidence used in the previous process. Furthermore, we utilized an improved prompt to generate 4.5K annotated instances for training the verb model.

The first author systematically traced and supervised the entire data generation process, ensuring a human-in-the-loop validation framework to maintain annotation quality and reliability. In addition, a representative sample of the generated data was evaluated using an LLM-as-a-judge approach (Li et al., 2025a) (See Appendix Table 11), employing OpenAI GPT-4o and DeepSeek-V3 as independent evaluators. The evaluation yielded consistently high scores across all dimensions. DeepSeek-V3 presents a average scores of 4.906 (Contextual Analysis), 4.898 (Justification Accuracy), 4.938 (Elimination Completeness), and 4.968 (Coherence), while GPT-4o shows 4.974, 4.962, 4.964, and 4.966, respectively. These results indicate strong agreement between evaluators and confirm the high quality of the generated dataset. Table 1 presents the statistics of the training data.

## 3.2. Fine-tuning strategies

The fine-tuning of the models was designed to address multiple objectives, allowing the evaluation of different performance criteria. Specifically, we assessed eight models to investigate whether low-parameter LLMs can effectively learn fine-tuning for accurate sense identification in the presence of ambiguous words. This proposed framework involves sequential tasks: (i) identifying the correct word sense without reasoning, (ii) performing neighbour-word analysis to identify the correct sense, and (iii) applying advanced reasoning that incorporates contextual understanding, justification of the correct sense, and refutation of incorrect senses to improve output quality.

We propose a novel **EAD framework** for tasks (ii) and (iii) consisting of three phases: *Exploration (E)*, *Analyzing (A)*, and *Disambiguating (D)*. In the *Exploration* phase, the framework collects sense inventories associated with the ambiguous word, including its interpretations and potential synonyms. The *Analyzing* phase emphasizes reasoning, which involves neighbour word analysis for task (ii) and a deeper evaluation of correct versus incorrect sense interpretations for task (iii). Finally, the *Disambiguating* phase consolidates the outcomes of the reasoning process to determine and retrieve the finalized sense ID. The models selected for the study include Gemma-2-2B (Team et al., 2024), Gemma-3-4B (Team et al., 2025), LLaMA-3.2-1B and LLaMA-3.2-3B (Dubey et al., 2024), Qwen-2.5-3B (Yang et al., 2025b), Qwen-3-4B (Yang et al., 2025a), SmoLM-3-3B and DeepSeek-Distill-Qwen1.5B (Liu et al., 2024). A systematic evaluation of these models was conducted to identify the best candidate to next phase.

### 3.2.1. Direct Sense Identification

To implement the first design objective "Given a word in a sentence, the LLM must generate the correct definition", we conducted a supervised evaluation of selected pre-trained language models on their ability to disambiguate the correct meaning of ambiguous words. For this phase, we utilized the FEWS dataset, which contains sentences annotated with target words and their corresponding senses. For instance, one training example includes the sentence: "He banked the plane sharply to avoid the storm," where the word "bank" is annotated with the sense "to tilt or incline an aircraft." The examples were formatted into instruction-response pairs, where the system prompt specifies the task, the input question provides the sentence and target word, and the output corresponds to the correct sense.

The pre-trained models were first evaluated via inference on these inputs using recommended hyperparameters to assess baseline disambiguation performance. Subsequently, we fine-tuned the baseline models using a supervised instruction fine-tuning approach, following the system prompt-input-output framework described in Subsection 3.3. Model performance was quantified using BERTScore, precision, recall, and F1 metrics to measure semantic alignment between predicted and reference senses. Initial experiments demonstrated that the Qwen-2.5 3B model exhibited promising learning capabilities after the initial training phase. To further investigate the impact of hyperparameters, additional experiments were conducted by varying the number of epochs and expanding the training dataset with additional SemCor records. These experiments allowed us to identify optimal hyperparameter configurations, which informed the subsequent phases of the study and ensured improved model generalization on the WSD.

### 3.2.2. Neighbour Words Analysis

The second design objective of "Given a word in a sentence and possible senses, the model reasons on the neighbouring context closely related to the disambiguation process to identify the correct sense" was developed in this phase. Inspired by Guzman-Olivares et al. (2025), we employed the fine-tuning design process following an EAD framework, giving full attention to the neighbouring words and following a CoT reasoning process for sense disambiguation. For the dataset creation, we implemented a context-extraction module using

a windowed semantic similarity approach. Specifically, each sentence was pre-processed to mark the ambiguous word using <WSD> tags, and spaCy was used to tokenize the preceding and following segments while filtering out stopwords and non-alphabetic tokens. From this, a fixed-size context window of up to 10 tokens on each side was extracted. The target word and its context tokens were then embedded using a sentence-transformer model, and cosine similarity scores were computed between the target embedding and each context token embedding. This allowed the contextual tokens to be ranked by semantic closeness to the target, from which the top-k (default k=5) most semantically relevant words were retained as features. While cosine similarity does not explicitly model syntactic dependencies, it serves as a salience-based heuristic to identify lexically influential neighbouring tokens within a constrained window. Importantly, the full sentence is retained during model fine-tuning, allowing the transformer architecture to capture long-range syntactic relations beyond the similarity-based feature selection. These context-sense pairs were then integrated into the dataset, ensuring that the training data explicitly encoded the most influential neighbouring words for disambiguation.

To exemplify the process, the sentence "After the match, the <WSD>bat</WSD> was placed carefully back into the player's bag" contains the target ambiguous word 'bat' which has 12 distinct noun senses according to the FEWS sense inventory. The preceding context tokens ['After', 'match'] and following context tokens ['placed', 'carefully', 'back', 'player', 'bag'] were extracted, and similarity scores were computed with 'bat'. The tokens were then ranked by their similarity values, and the top-k most semantically related neighbours were selected. In this example, the high similarity of 'match', 'player', and 'bag' to 'bat' strongly indicates the sports equipment sense, as opposed to the flying mammal or old woman senses. These ranked neighbours serve as semantically aligned cues from the local context, guiding the disambiguation process toward the correct interpretation.

We applied this process to annotate 101K FEWS and 226K SemCor data records, structuring the output in CoT reasoning, then used this dataset to fine-tune the models. Our experiments revealed that this approach was more efficient than standard fine-tuning, yielding promising results for disambiguation. However, we observed that LLaMA 1B, LLaMA 3B, and DeepSeek-distill-Qwen 1.5B did not acquire the expected level of reasoning capability. The other models demonstrated strong reasoning performance, showing robust generalization and effective disambiguation even for unseen sentences.

#### 3.2.3. Advanced Reasoning

This phase was designed to achieve the goal: "Given a word in a sentence and its possible senses, the model reasons why a particular sense is correct and why the others are not," to determine the correct meaning. The main objective of this process is to disambiguate meaning through a deep analysis of each sense in relation to the sentence requiring disambiguation. Due to the lack of an appropriate reasoning-specific dataset, we used a large-parameter open source LLM with a carefully designed prompt. We used a dataset of 10K samples which were randomly selected from the FEWS dataset to ensure coverage across all POS tags.

We discovered that three elements are critical to effective reasoning in sense disambiguation: (i) contextual analysis, (ii) justification of the correct sense, and (iii) systematic elimination of incorrect senses. Our design was inspired by Huang et al. (2020), who reported the importance of using context-gloss pairs to train models not only to identify why a sense ID is valid but also to reject invalid ones. Adopting the setup described in section 3.3, we fine-tuned the selected model and achieved comparable results despite the small sample size, demonstrating improved reasoning performance for disambiguation. This process aligns with the EAD framework, where exploration gathers candidate senses, analysis provides reasoning for sense selection, and disambiguation yields the final sense ID.

After the initial reasoning process, we observed limitations in verb disambiguation. To address this challenge, we incorporated syntactic evidence into the reasoning process, where a verb's morphosyntax (tense, aspect, voice), immediate dependents (subjects, objects, complements), key function words (auxiliaries, particles, prepositions), and relevant dependency or constituent patterns were explicitly included in the prompt to constrain its meaning. To evaluate the effect of these modifications, we trained and tested Qwen-3 and Gemma-3 models exclusively on newly constructed training data tailored for this purpose. The evaluation was conducted using the FEWS Few-shot Development set with verb subset, consisting of 2.2K samples, ensuring the assessment was focused on verb disambiguation. Moreover, we employed a hybrid training approach that combined reasoning with neighbour, syntactic, and semantic analysis, and the results were recorded to validate the improvements in learning and disambiguation accuracy.

#### 3.2.4. Ablation Study

The experimental findings showed that reasoning-based approaches are well-suited for handling unseen data. To evaluate these findings further, we examine the generalizability of such models to dif-

ferent datasets without any additional fine-tuning.

Thus, we employed the recently released "Fool Me If You Can" dataset (Ballout et al., 2024), an adversarial benchmark designed to investigate the robustness of language models in WSD. We selected this dataset for the ablation study due to the varied contextual challenges. Specifically, we use the FEWS sense-mapped version retrieved from prior work (Sumanathilaka et al., 2025b) [5]. The dataset consists of four subsets: a baseline with context-appropriate meanings similar to training data, for standard WSD evaluation; a version adding adjectives that reinforce intended meanings; another replacing these with adjectives linked to the opposite sense; and newly crafted natural sentences where context contradicts the expected meaning, creating harder disambiguation scenarios. Fine-tuning on the "Fool Me If You Can" training data was intentionally avoided to ensure that the evaluation measured the robustness and generalization capabilities of the current reasoning models when exposed to unseen data from a different dataset.

To further probe these capabilities, we also evaluated the models on two additional challenging benchmarks: hardEn (Maru et al., 2022), which comprises 476 test cases that cannot be solved by existing state-of-the-art WSD models, and 42D, a specialized dataset containing 370 records with rare and domain-specific senses as defined by BabelNet (Navigli and Ponzetto, 2010).

This design choice allows us to assess whether the models can maintain disambiguation performance under domain shift and adversarial context, rather than benefiting from dataset-specific adaptation (see Section 4.2 for results).

### 3.3. Study Setup

For model development, supervised fine-tuning (SFT) was employed to adapt a pre-trained LLM to WSD task. The baseline models were fine-tuned using Low-Rank Adaptation (LoRA) to enable efficient training while reducing computational overhead. Training data were pre-processed into a desired chat-style prompt–response format, with each example containing a context sentence and the ambiguous target word to be disambiguated.

Fine-tuning was conducted using Hugging Face transformers and trl libraries, with a custom prompt formatting function to standardize inputs. The datasets were tokenized using the model's native tokenizer, and training hyperparameters: batch size $= 4$, gradient accumulation $= 8$ steps, learning rate $= 2 \times 10^{-4}$. We used AdamW optimizer and linear learning rate scheduler with a random seed of 3407 as recommended by prior work (Picard, 2023). The fine-tuning used standard supervised causal language modelling, with cross-entropy loss backpropagated through the gold output tokens.

All experiments were performed on an NVIDIA A100-PCIE-40GB GPU, ensuring sufficient memory bandwidth for efficient fine-tuning of the model without quantization in the final setup. The resulting LoRA adapters and tokenizer were saved locally and uploaded to the Hugging Face Hub for reproducibility and deployment, which can be accessed at `https://huggingface.co/deshanksuman`.

## 4. Results and Discussion

In the initial phase, we evaluated eight different low-parameter LLMs to assess their performance in disambiguating a word within a given sentence. Inspired by Sumanathilaka et al. (2024a), we evaluate using a subset of the FEWS test dataset containing 1,050 cases distributed across nouns, verbs, and adjectives in a 4:3:3 ratio and with 50 adverb examples, to ensure coverage across parts-of-speech.

Performance was measured using BERTScore, computed with the distilled base uncased models. BERTScore was chosen because the evaluation focused on comparing the semantic similarity between the expected meaning and the model's predicted output, rather than exact token matching. As the initial experiments were not sense-mapped, a meaning-based metric provided a more informative estimate of model performance.

We applied an SFT approach without explicit reasoning steps to examine how well the models could learn the senses of ambiguous words. Fine-tuning revealed clear improvements in sense identification and disambiguation for the Qwen and LLaMA models (see Table 4). Gemma-3-4B showed a performance drop in the sense identification setting. However, this was not due to poor disambiguation; rather, the model tended to generate longer, more detailed sense explanations rather than concise labels. Because evaluation relied on BERTScore against short reference descriptions, the resulting outputs were more verbose, leading to lower similarity scores. Notably, this issue did not appear in the sense ID prediction setting with predefined candidates.

Given the better performance of Qwen-2.5-3B, we further examined the effect of the number of training epochs. We found that increasing the number of epochs from one to five did not yield significant improvements in performance for F1 score (1:0.568, 3: 0.552, 5:0.571). McNemar tests (McNemar, 1947) did not reveal any statistically significant differences. Thus, we limited fine-tuning to two epochs to avoid unnecessary computation overhead.

[5] https://github.com/Sumanathilaka/FOOL-ME-IF-YOU-CAN-dataset-Meets-FEWS-sense-Tags

Table 2: Bert Score based on distilled base uncased against the base model vs finetuned model. The best scores are in bold. * Deepseek-distill-Qwen 1.5B

| Models | Base Inference | | | Instructional Fine-Tuning | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score |
| Gemma 2 2B | 0.7720 | 0.7341 | 0.7515 | 0.7870 | 0.7781 | 0.7814 |
| Gemma 3 4B | **0.7806** | 0.7302 | 0.7534 | 0.7741 | 0.7214 | 0.7458 |
| Llama 3.2 1B | 0.6439 | 0.7503 | 0.6923 | 0.7919 | 0.7474 | 0.7681 |
| Llama 3.2 3B | 0.7616 | 0.7330 | 0.7430 | **0.8079** | 0.7567 | 0.7804 |
| SmolLM 3 | 0.7687 | **0.7656** | **0.7662** | 0.7884 | 0.7636 | 0.7749 |
| Qwen 2.5 3B | 0.7551 | 0.6773 | 0.7130 | 0.8126 | **0.8018** | **0.8061** |
| Qwen 3 4B | 0.7619 | 0.7470 | 0.7531 | **0.8128** | 0.7970 | 0.8036 |
| Deepseek 1.5B* | 0.7248 | 0.7193 | 0.7212 | 0.7140 | 0.7491 | 0.7294 |

### 4.1. Reasoning Experiments

In the initial experiments of the second phase, we fine-tuned the eight models to evaluate their ability to learn reasoning using neighbouring word analysis. All models were initially trained for 1 epoch to assess performance and selected models were extended to 2 epochs based on performance. Notably, Llama 1B, 3B, and Deepseek-distill-Qwen 1.5B did not learn reasoning effectively and behaved primarily as text predictors, failing to follow the instructions given during inference.

As presented in Table 3, fine-tuning significantly improves performance over the baseline. For instance, Qwen-3-4B and Gemma-3-4B models showed substantial gains after 2 epochs, achieving the highest overall scores of 0.738 and 0.752, respectively. Among the parts of speech, Gemma-3-4B achieved the highest score in Adverbs (0.80) after 1 epoch, while Qwen-3-4B showed the best improvement in Adjectives (0.75) after 2 epochs. These observations indicate that fine-tuning leads to consistent improvements across all categories, confirming that small models with reasoning-based fine-tuning benefit more from extended training compared to baselines.

To evaluate the effectiveness of different reasoning strategies, we benchmark the CoT reasoning-based approach against the few-shot reasoning based approach and the advanced reasoning with correct and incorrect sense analysis. For the few-shot reasoning based approach, we employed the same fine-tuned models; however, during inference, we provided additional few-shot examples to further emphasize understanding of the target senses, in line with evidence that supplementary contextual examples could improve performance (Yang et al., 2024). Each sense was accompanied by two illustrative examples that demonstrated its meaning and interpretation within a specific context. For evaluation, the dynamic few-shot examples were retrieved from a pre-constructed knowledge base.

Table 4 indicates that the zero-shot CoT reasoning approach generally outperforms the few-shot reasoning approach in terms of overall accuracy, particularly for the Gemma-3-4B model fine-tuned over two epochs (0.75). Interestingly, this zero-shot CoT configuration also achieves the highest noun (0.81) and verb (0.71) scores across all tested settings. While the few-shot approach provides additional contextual cues through example cases, it does not consistently translate into higher performance, suggesting that the models are already able to leverage their fine-tuned reasoning capabilities without supplementary examples. The advanced reasoning with correct and incorrect sense analysis approach performs competitively, with Qwen-3-4B reaching strong adverb performance (0.80) and stable results across categories, but still not surpassing the zero-shot CoT reasoning in overall accuracy. Importantly, the advanced reasoning approach used only 10% of the training data compared to the CoT reasoning approach, yet produced comparable results. This trend underscores the robustness of the CoT reasoning paradigm, even in the absence of explicit few-shot demonstrations, and highlights its ability to generalize reasoning patterns effectively across different POS categories.

To gain a broader understanding of the performance of the proposed reasoning-based approach compared to low-parameter models, we evaluated on the FEWS test set, which contains both few-shot and zero-shot datasets. Notably, the few-shot setting included examples whose senses do appear in the training set, but only in limited numbers, whereas the zero-shot setting contained examples of senses not encountered during training. We benchmarked our results against current state-of-the-art models to assess the effectiveness of our approach. Thus, we compared against MFS, Lesk, SemEq (Yao et al., 2021), ESR (Song et al., 2021), RTWE (Zhang et al., 2023), and GlossGPT (Sumanathilaka et al., 2025b) (see Table 5).

The proposed models (Qwen & Gemma) show competitive performance despite having significantly fewer parameters than most competitor systems. With the Few-shot set, both models outperform traditional baselines (i.e. MFS & Lesk) by a considerable margin and achieve close to state-of-the-art results. Notably, Qwen achieves 76.52 and

Table 3: Fine Tuning with CoT based reasoning with Neighbour words analysis. The results are presented in the F1 score.

| Models | Noun | Verb | Adjective | Adverb | Overall |
|---|---|---|---|---|---|
| **Fine Tuned with CoT based Neighbour words analysis Approach** | | | | | |
| SmolLM 3B (1 epoch) | 0.52 | 0.35 | 0.50 | 0.54 | 0.47 |
| Gemma-2 2B (1 epoch) | 0.70 | 0.61 | 0.68 | 0.66 | 0.67 |
| Qwen 2.5 3B (1 epoch) | 0.71 | 0.64 | 0.71 | 0.74 | 0.69 |
| Gemma 3 4B (1 epoch) | 0.79 | 0.65 | 0.70 | **0.80** | 0.72 |
| Qwen 3 4B (2 epochs) | 0.79 | 0.67 | **0.75** | 0.68 | 0.74 |
| Gemma 3 4B (2 epochs) | **0.81** | **0.71** | 0.72 | 0.76 | **0.75** |
| **Current Baseline** | | | | | |
| Qwen 3 4B | 0.69 | 0.54 | 0.58 | 0.48 | 0.61 |
| Gemma 3 4B | 0.62 | 0.48 | 0.52 | 0.48 | 0.54 |
| Gemma 7B | 0.49 | 0.41 | 0.51 | 0.46 | 0.47 |
| Mixtral 7B | 0.43 | 0.32 | 0.46 | 0.42 | 0.41 |
| Yi - 34B | 0.65 | 0.51 | 0.57 | 0.52 | 0.58 |
| GPT 4o-mini | 0.37 | 0.30 | 0.31 | 0.32 | 0.33 |

Table 4: F1 Score for benchmarking Different Reasoning Strategies.

| Models | Noun | Verb | Adjective | Adverb | Overall |
|---|---|---|---|---|---|
| **Fine Tuned with COT based Neighbour words analysis Approach (Zero shot)** | | | | | |
| Qwen 3 4B (2 epochs) | 0.79 | 0.67 | **0.75** | 0.68 | 0.74$^{**}$ |
| Gemma 3 4B (2 epochs) | **0.81** | **0.71** | 0.72 | **0.76** | **0.75**$^{**}$ |
| **Fine Tuned with COT based Neighbour words analysis Approach (Few shot)** | | | | | |
| Qwen 3 4B (2 epochs) | 0.74 | 0.62 | 0.68 | 0.64 | 0.68 |
| Gemma 3 4B (2 epochs) | **0.78** | **0.68** | **0.70** | **0.78** | **0.72** |
| **Advanced Reasoning with Correct and Incorrect Sense Analysis** | | | | | |
| Gemma 3 4B (2 epochs) | 0.76 | 0.60 | 0.66 | 0.64 | 0.68 |
| Qwen 2.5 3B (2 epochs) | 0.74 | 0.61 | 0.68 | 0.66 | 0.68 |
| Qwen 3 4B (2 epochs) | **0.76** | **0.67** | **0.70** | **0.80** | **0.72** |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Gemma 75.68, surpassing $\text{RTWE}_\text{L}$ and approaching GlossGPT's performance. In the more challenging Zero-shot set, the models maintain strong accuracy (72.66 and 71.86), outperforming most baselines, showing robust generalization. The results show the effectiveness of our reasoning-based fine-tuning approach, even using compact models.

From Table 4, we observed that verb disambiguation remains challenging for all models. We explored different fine-tuning strategies to test model reasoning improvements (see setup in section 3.2.3). However, verb-specific models still struggle to disambiguate verbs using syntactic clues, even when fine-tuned or combined in a hybrid model that incorporates both reasoning and neighbour analysis with syntactic and semantic features. Table 6 shows that adding extra reasoning signals still fails due to being interpreted as noise.

## 4.2. Ablation Study

The ablation study aimed to evaluate the robustness and generalizability of the trained module by applying CoT based reasoning combined with Neighbour Words Analysis to new, unseen data of two diverse use cases.

Table 7 presents a detailed comparison of our proposed approach against several state-of-the-art WSD systems on the challenging hard WSD benchmarks introduced by Maru et al. (2022). Our method, leveraging neighbour word analysis combined with Finetuned Qwen and Gemma, demonstrates superior performance, with Qwen-4B achieving the highest F1 scores across both 42D and hardEN datasets. Notably, Qwen-4B attains an F1 score of 78.48 on the hardN benchmark, substantially outperforming existing models, while Gemma also shows strong competitive results.

Table 8 shows that the model adapts effectively to unseen data with the Fool dataset, maintaining comparable performance across Sets 1-3. This underscores the extent to which the model can handle novel scenarios and maintain robust reasoning capabilities. Importantly, in the more challenging Set 4, containing realistic sentences where contextual cues deliberately contradict the expected meaning of a homonym, the model reached state-of-the-art performance, even for the small LLM class. This demonstrates the model's capacity to address difficult disambiguation tasks that require nuanced contextual understanding. When compared to the

Table 5: F1 score for CoT based Neighbour words analysis against current techniques. FEWS test data used for evaluation. $_b$: Base, $_L$: Large, Our approach in italic (4B models)

| Dataset | MFS | Lesk | SemEq$_b$ | SemEq$_L$ | ESR$_b$ | ESR$_L$ | RTWE$_L$ | *Qwen* | *Gemma* | GlossGPT |
|---|---|---|---|---|---|---|---|---|---|---|
| Fewshot | 51.5 | 40.9 | 80.1 | 82.3 | 77.8 | 83.4 | 78.4 | *76.52* | *75.68* | **90.7** |
| Zeroshot | - | 39.0 | 70.2 | 72.2 | 71.6 | 75.8 | 69.9 | *72.66* | *71.86* | **79.5** |

Table 6: F1 score of the VERB task compared using different reasoning methods. Hybrid:Neighbour Word Analysis with Syntactic & Semantic Analysis, A: Analysis

| Model | Neighbour Word A. | Semantic A. | Syntactic & Semantic A. | Hybrid |
|---|---|---|---|---|
| Qwen3-4B | **0.679**$^{*}$ | 0.662 | 0.677 | 0.602 |
| Gemma3-4B | **0.658**$^{**}$ | 0.601 | 0.628 | 0.544 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: Performance comparison (F1 score) against existing WSD models. The compared models are ARES (Scarlini et al., 2020),BEM (Blevins and Zettlemoyer, 2020), ESC (Barba et al., 2021a), EWS (Bevilacqua et al., 2020b), GEN (Bevilacqua et al., 2020a), SYN (Scozzafava et al., 2020)),CSC (Barba et al., 2021b) and SandWiCH (Guzman-Olivares et al., 2025). Our best-performing approach, Neighbour words analysis with Finetuned Qwen and Gemma, is reported in Italic.

| Dataset | ARES | BEM | ESC | EWS | GEN | SYN | CSC | SandWiCH | *Gemma* | *Qwen* |
|---|---|---|---|---|---|---|---|---|---|---|
| 42D | 41.8 | 53.2 | 58.9 | 43.9 | 50.2 | 32.8 | 56.6 | 77.1 | 64.43 | **78.48** |
| hardEN | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.35 | 53.4 | 40.71 | **54.19** |

Table 8: F1 score for "Fool me if you can" dataset for binary classification of sense ID. *Our approach is not fine-tuned with training data from "Fool me if you can" dataset.

| Models | # Parameters | Set 1 | Set 2 | Set 3 | Set 4 |
|---|---|---|---|---|---|
| Roberta-base | 125M | 0.945 | 0.969 | 0.888 | 0.715 |
| Bert-large | 340M | 0.970 | 0.978 | 0.874 | 0.689 |
| T5-large | 770M | 0.984 | 0.987 | 0.896 | 0.691 |
| FLAN-T5-large | 780M | 0.948 | 0.953 | 0.852 | 0.663 |
| T5-xl | 3B | **0.991** | 0.992 | **0.907** | 0.710 |
| FLAN-T5-xl | 3B | 0.955 | 0.958 | 0.881 | 0.718 |
| Mixtral 7bx8 | 7B | 0.987 | **0.993** | 0.820 | 0.714 |
| Llama3 8B | 8B | 0.986 | 0.990 | 0.790 | 0.687 |
| *Gemma 3 (Reasoning)** | 4B | 0.966 | 0.969 | 0.812 | 0.811 |
| *Qwen 3 (Reasoning)** | 4B | 0.970 | 0.972 | 0.847 | **0.852** |

results reported in the original paper (Ballout et al., 2024), our approach not only outperformed GPT-3.5-Turbo but also revealed that the key driver of superior performance is not merely model size, but the inclusion of a well-structured reasoning process. This reinforces the importance of targeted reasoning strategies for achieving strong generalization and adaptability in language models. This is an important finding because it challenges the common assumption of larger sized models for improved performance (Yae et al., 2025). Instead, it reveals that carefully designed reasoning strategies like *CoT combined with Neighbour Words Analysis* can yield state-of-the-art results even with smaller models.

## 5. Conclusion and Future Work

Our study shows that low-parameter LLMs can achieve competitive and state-of-the-art performance in WSD when equipped with reasoning-based fine-tuning strategies. Through evaluations with eight <4B parameter models, we showed that CoT reasoning combined with neighbour-word analysis enables strong sense prediction even in zero-shot and domain-shift settings. Gemma-3-4B and Qwen-3-4B models consistently outperformed medium-parameter baselines and were comparable to larger models (e.g., GPT-3.5-Turbo), while maintaining robustness across adversarial datasets and hard WSD datasets.

Importantly, the proposed approaches were effective in modest computational settings, with the advanced reasoning strategy achieving competitive accuracy using only 10% of the training data compared to the CoT method. These findings support the fact that reasoning quality is a critical determinant of LLM-based WSD performance. Future work should extend this approach to multilingual settings, including low-resource languages, to assess its adaptability across linguistic contexts.

## Limitations

The scope of this study is limited to eight models from different vendors, all of which have fewer than 4B parameters. While further exploration of mid-sized models may yield improved performance, the primary objective, demonstrating performance gains through enhanced reasoning has been successfully achieved and remains adaptable. The current study focuses exclusively on English WSD; however, the approach can be extended to multilingual or cross-lingual settings in future work. Due to computational constraints, training iterations were restricted to two epochs, and only a sample of the generated data was validated owing to limited resource availability. Nevertheless, the evaluated samples produced promising results consistent with the expectations for this study.

## Acknowledgements

We acknowledge the support of the Super computing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government. Hough's work is supported by the EPSRC grant EP/X009343/1 'FLUIDITY'.

## Ethics consideration

This paper has been conducted in compliance with Swansea University's ethical standards.

## Data & Code availability

The fine-tuned LoRA adapters used in this study are publicly available on Hugging Face[6], enabling reproducibility and further research. In addition, the full training, inference, and evaluation pipeline has been released as open-source code on GitHub[7].

## Bibliographical References

Reem Abdel-Salam. 2024. rematchka at arabic-nlu2024: Evaluating large language models for arabic word sense and location sense disambiguation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 383–392.

Eneko Agirre, Oier López De Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Boshra F Zopon Al-Bayaty and Shashank Joshi. 2016. Comparative analysis between naïve bayes algorithm and decision tree to solve wsd using empirical approach. *Lecture Notes on Software Engineering*, 4(1):82.

Joseph Bamidele Awotunde. 2025. Word sense disambiguation in biomedical applications. In *Mining Biomedical Text, Images and Visual Features for Information Retrieval*, pages 587–605. Elsevier.

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. 2024. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Mohamad Ballout, Anne Dedert, Nohayr Abdelmoneim, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. 2024. Fool me if you can! an adversarial dataset to investigate the robustness of lms in word sense disambiguation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5042–5059.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning WSD with Extractive Sense Comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models. *arXiv preprint arXiv:2503.08662*.

Michele Bevilacqua, Marco Maru, Roberto Navigli, et al. 2020a. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference*

[6] https://huggingface.co/deshanksuman

[7] https://github.com/Sumanathilaka/An-EAD-Reasoning-Framework-for-WSD-with-Low-Parameter-LLMs

*on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221.

Michele Bevilacqua, Roberto Navigli, et al. 2020b. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Melissa Bowerman. 1978. The acquisition of word meaning: An investigation into some current conflicts. In *The development of communication*, pages 263–287. Wiley.

Samuel Cahyawijaya, Ruochen Zhang, Holy Lovenia, Jan Christian Blaise Cruz, Hiroki Nomoto, and Alham Fikri Aji. 2024. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word sense. *arXiv preprint arXiv:2410.21573*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Robert David, Anna Kernerman, Ilan Kernerman, Nicolas Ferranti, and Assaf Siani. 2024. Multilingual word sense disambiguation for semantic annotations: Fusing knowledge graphs, lexical resources, and large language models. In *Proceedings of CEUR Workshop (RAGE-KG 2024)*, pages 16–22.

José Marcio Duarte, Samuel Sousa, Evangelos Milios, and Lilian Berton. 2021. Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. *Information Sciences*, 570:278–297.

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 e-prints*, pages arXiv–2407.

Gurinder Pal Singh Gosal. 2015. A naive bayes approach for word sense disambiguation. *International Journal*, 5(7).

Roksana Goworek, Harpal Karlcut, Muhammad Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Purighella, et al. 2025. Senwich: Sense-annotation of low-resource languages for wic using hybrid methods. *arXiv preprint arXiv:2505.23714*.

Daniel Guzman-Olivares, Lara Quijano-Sanchez, and Federico Liberatore. 2025. Sandwich: Semantical analysis of neighbours for disambiguating words in context ad hoc. *arXiv preprint arXiv:2503.05958*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2020. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. ArXiv:1908.07245 [cs].

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2023. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. *arXiv preprint arXiv:2304.13803*.

SG Kolte and SG Bhirud. 2009. Exploiting links in wordnet hierarchy for word sense disambiguation of nouns. In *Proceedings of the International Conference on Advances in Computing, Communication and Control*, pages 20–25.

Neeraja Koppula, K. Srinivasa Rao, and B. VeeraSekharReddy. 2021. Word Sense Disambiguation Using Context Dependent Methods. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1582–1590, Tirunelveli, India. IEEE.

Sunjae Kwon, Dongsuk Oh, and Youngjoong Ko. 2021. Word sense disambiguation based on context selection using knowledge-based word similarity. *Information Processing & Management*, 58(4):102551.

Rim Laatar, Chafik Aloulou, and Lamia Hadrich Belguith. 2023. Evaluation of Stacked Embeddings for Arabic Word Sense Disambiguation. *Computación y Sistemas*, 27(2).

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych.

2023. Contextual embeddings for ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19.

Cuong Anh Le and Akira Shimazu. 2004. High wsd accuracy using naive bayesian classifier with rich features. In *Proceedings of the 18th Pacific Asia conference on language, information and computation*, pages 105–114.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025b. Exploring the impact of temperature on large language models: Hot or cold? *arXiv preprint arXiv:2506.07295*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating Glosses into Neural Word Sense Disambiguation. ArXiv:1805.08028 [cs].

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the Hard Core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33885–33904.

Fanqing Meng. 2022. Word sense disambiguation based on graph and knowledge base. In *4th EAI international conference on robotic sensor networks*, pages 31–41. Springer.

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.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in ChatGPT. ArXiv:2302.06426 [cs].

Martha Palmer, Hwee Tou Ng, and Hoa Trang Dang. 2006. Evaluation of wsd systems. *Word Sense Disambiguation: Algorithms and Applications*, pages 75–106.

David Picard. 2023. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16339–16347.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Zhengfei Ren, Annalina Caputo, and Gareth Jones. 2024. A few-shot learning approach for lexical semantic change detection using gpt-4. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 187–192.

Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre, and German Rigau. 2023. What do language models know about word senses? zero-shot WSD with language models and domain inventories. In *Proceedings of the 12th Global Wordnet Conference*, pages 331–342, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word SenseDisambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020)*. Association for Computational Linguistics.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Tajinder Singh, Madhu Kumari, and Daya Sagar Gupta. 2024. Context-based persuasion analysis of sentiment polarity disambiguation in social media text streams. *New Generation Computing*, 42(4):497–531.

Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved Word Sense Disambiguation with Enhanced Sense Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4311–4320, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deshan Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024a. Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques. In *2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC)*, pages 204–209. IEEE.

Deshan Sumanathilaka, Nicholas Micallef, and Julian Hough. 2025a. Exploring the impact of temperature on large language models: A case study for classification task based on word sense disambiguation. In *2025 7th International Conference on Natural Language Processing (ICNLP)*, pages 178–182.

Deshan Sumanathilaka, Nicholas Micallef, and Julian Hough. 2025b. GlossGPT: GPT for Word Sense Disambiguation using Few-shot Chain-of-Thought Prompting. *Procedia Computer Science*.

Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024b. Can LLMs assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2025c. Prompt balance matters: Understanding how imbalanced few-shot learning affects multilingual sense disambiguation in LLMs. In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 7–15, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

Ming Wang and Yinglin Wang. 2020. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Jung H Yae, Nolan C Skelly, Neil C Ranly, and Phillip M LaCasse. 2024. Leveraging large lan-

guage models for word sense disambiguation. *Neural Computing and Applications*, pages 1–18.

Jung H Yae, Nolan C Skelly, Neil C Ranly, and Phillip M LaCasse. 2025. Leveraging large language models for word sense disambiguation. *Neural Computing and Applications*, 37(6):4093–4110.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025b. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024. Multiple sources are better than one: Incorporating external knowledge in low-resource glossing. *arXiv preprint arXiv:2406.11085*.

Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. 2021. Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories. ArXiv:2110.14091 [cs].

Xuefeng Zhang, Richong Zhang, Xiaoyang Li, Fanshuang Kong, Junfan Chen, Samuel Mensah, and Yongyi Mao. 2023. Word Sense Disambiguation by Refining Target Word Embedding. In *Proceedings of the ACM Web Conference 2023*, pages 1405–1414, Austin TX USA. ACM.

# 6. Appendix

Table 9: Prompt used to elicit rationales for sense selection during fine-tuning.

You are a helpful assistant tasked with simulating the human thinking process for Word Sense Disambiguation (WSD). Given a sentence containing an ambiguous word, along with the expected sense ID, your goal is to generate a detailed and logical reasoning process as a human would.
Your response should include:

- Contextual Analysis – Explain how the surrounding context in the sentence helps determine the correct meaning of the ambiguous word.
- Justification of the Correct Sense ID – Provide a clear explanation of why the expected sense ID is appropriate in this context.
- Elimination of Incorrect Senses – Briefly describe why the other possible sense IDs do not fit the given context.

Be thorough, logical, and emulate how a human would naturally think through the disambiguation. Do not include any additional information or instructions outside of the reasoning process.

Table 10: Prompt used to elicit rationales for Verb sense selection during extended study.

You are a helpful assistant tasked with simulating the human thinking process for Word Sense Disambiguation (WSD).
Given a sentence containing an ambiguous word, along with the expected sense ID, your goal is to generate a detailed and logical reasoning process as a human would.
The given ambiguous word is a VERB.
Your response should include:

1. Syntactic Evidence — Summarize the verb's morphosyntax (tense/aspect/voice), immediate dependents (subject, objects, complements), key function words (auxiliaries, particles, prepositions), and any relevant dependency/constituent patterns that constrain its meaning.
2. Semantic Evidence — Describe selectional preferences and semantic roles implied by the verb, plausible paraphrases, collocations, and context/topic cues (entities, events) that support a specific sense.
3. Decision — State the chosen sense ID and give a clear justification that ties the syntactic and semantic cues to that sense.
4. Elimination of Alternatives — Briefly state why other possible sense IDs do not fit given the observed syntax and semantics.

Be thorough, logical, and emulate how a human would naturally think through the disambiguation. Do not include any additional information or instructions outside of the reasoning process.

Table 11: Prompt used for LLM-as-a-Judge evaluation of WSD reasoning quality

You are an expert evaluator assessing the quality of Word Sense Disambiguation (WSD) reasoning generated by an AI system.

**Task Context:**

- Original sentence, ambiguous word, and sense definitions: {input_text}
- Correct sense ID: {senseid}
- Generated Reasoning to evaluate: {reasoning}

**Evaluation Instructions:**

Evaluate the generated reasoning on the following four dimensions using a 1-5 scale:

1. **Contextual Analysis Quality**: Does the reasoning identify and explain relevant contextual clues effectively?
    - 5: Comprehensively identifies all relevant contextual clues and explains their disambiguation role
    - 4: Identifies most key contextual elements with clear explanations
    - 3: Identifies some context but misses important clues or lacks depth
    - 2: Minimal contextual analysis with superficial observations
    - 1: Incorrect or irrelevant contextual analysis
2. **Sense ID Justification Accuracy**: Is the justification for the correct sense logically sound and semantically accurate?
    - 5: Provides precise, logically sound justification with clear semantic connections
    - 4: Good justification with minor gaps in reasoning
    - 3: Adequate justification but lacks depth or contains minor logical errors
    - 2: Weak justification with significant logical gaps
    - 1: Incorrect or nonsensical justification
3. **Elimination Reasoning Completeness**: Does it systematically eliminate alternative senses with clear reasoning?
    - 5: Systematically eliminates all alternative senses with clear reasoning
    - 4: Eliminates most alternatives with good reasoning
    - 3: Partial elimination with some reasoning gaps
    - 2: Minimal elimination or weak reasoning
    - 1: No elimination or incorrect reasoning
4. **Overall Coherence and Human-likeness**: Does the reasoning flow naturally like human thinking?
    - 5: Reads like natural human reasoning, logically structured and comprehensive
    - 4: Good flow with minor artificiality
    - 3: Adequate but somewhat mechanical or repetitive
    - 2: Disjointed or overly formulaic
    - 1: Incoherent or completely artificial

**Output Format (JSON):**

```
{
  "contextual_analysis_score": 1-5,
  "justification_accuracy_score": 1-5,
  "elimination_completeness_score": 1-5,
  "coherence_score": 1-5
}
```