Zero-Shot and Fine-Tuned Evaluation of Generative LLMs for Arabic Word Sense Disambiguation

Yossra Noureldien, Abdelrazig Mohamed, Farah Attallah
University of Khartoum
{yossra.noureldien, abdelrazig.mohamed, farah.hassan}@uofk.edu

Abstract

Arabic presents unique challenges for sense-level language understanding due to its rich morphology and semantic ambiguity. This paper benchmarks large generative language models (LLMs) for Arabic Word Sense Disambiguation (WSD) under both zero-shot and fine-tuning conditions. We evaluate one proprietary model (GPT-4o) and three open-source models (LLaMA 3.1-8B, Qwen 2.5-7B, and Gemma 2-9B) on two publicly available datasets. In zero-shot settings, GPT-4o achieved the highest overall performance, with comparable results across both datasets, reaching 79% accuracy and an average macro-F1 score of 66.08%. Fine-tuning, however, notably elevated all open models beyond GPT-4o's zero-shot results. Qwen achieved the top scores on one dataset, with an accuracy of 90.77% and a macro-F1 score of 83.98%, while LLaMA scored highest on the other, reaching an accuracy of 88.51% and a macro-F1 score of 69.41%. These findings demonstrate that parameter-efficient supervised adaptation can close much of the performance gap and establish strong, reproducible baselines for Arabic WSD using open-source, relatively medium-sized models. Full code is publicly available.[1]

1 Introduction

Word Sense Disambiguation (WSD) is a core problem in Natural Language Processing (NLP) that involves determining which sense of a word is intended within a particular context. This task is especially challenging due to semantic polysemy, where individual words can convey multiple meanings depending on their context of use. Arabic, in particular, significantly amplifies this complexity due to its rich morphological structure and substantial polysemy (Al-Hajj and Jarrar, 2021; Kaddoura and Nassar, 2024b).

[1] https://github.com/Yossranour1996/Arabic-WSD-LLM

فَقَدَ نَفْسَ عَزِيزٍ عَلَيْهِ (1)

أَصْدَرَ القَاضِي نَفْسَ الحُكْمِ (2)

A representative example is the word (نَفْس, *nafs*), which has different meanings depending on the context. In sentence (1), (*faqada nafs ʿazīzin ʿalayhi*), meaning (*he lost a dear soul*), the word refers to (soul). In sentence (2), (*aṣdara al-qāḍī nafs al-ḥukm*), meaning (*the judge issued the same ruling*), it means (same).

Furthermore, omitting diacritics in written Arabic exacerbates ambiguity, complicating the task of accurate disambiguation (Alqahtani et al., 2019).

Before the advent of modern Artificial Intelligence (AI) methods, traditional approaches dominated WSD tasks. These older methods primarily involved rule-based strategies utilizing lexical databases and glossaries, alongside statistical and dictionary-based approaches (Abeysiriwardana and Sumanathilaka, 2024; Eid et al., 2010). Although foundational, these traditional methodologies exhibited limitations in scalability and contextual adaptability.

Recent advancements in NLP have introduced powerful Large Language Models (LLMs) that significantly enhance the ability to address semantic tasks through the use of contextualized representations. Encoder-based models, such as BERT (Devlin et al., 2019), have demonstrated high effectiveness in various language understanding tasks through supervised fine-tuning on labeled data. In Arabic, adaptations such as AraBERT (Antoun et al., 2020) and CAMeLBERT (Inoue et al., 2021) have enabled the capture of Arabic linguistic features more effectively.

On the generative side, autoregressive decoder-based models such as the GPT series (Radford et al., 2018), and newer multilingual and Arabic-capable models like LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023), Jais and Jais-chat (Sen-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 298–305

gupta et al., 2023), and ALLaM (Bari et al., 2024) have opened the door for zero-shot, few-shot, and fine-tuning–based learning approaches. Unlike encoder-based architectures, these generative models operate by predicting the next token in a sequence, making them well-suited for prompt-based inference and instruction-following settings. This architectural distinction underpins differences in how each family of models performs disambiguation, offering complementary strengths for WSD evaluation.

Despite these advancements, the potential of generative LLMs for Arabic WSD remains not well explored. In this paper, we provide the following contributions:

- **We analyze** available Arabic WSD datasets and identify those most suitable for evaluation.
- **We evaluate** generative LLMs under zero-shot and fine-tuned settings, assessing their effectiveness in Arabic sense disambiguation.

2 Related Work

The recent rise of Pre-trained Language Models (PLMs) has significantly advanced NLP, leading to extensive efforts to benchmark their effectiveness across diverse linguistic contexts and a wide range of Arabic NLP tasks.

For instance, ORCA (Elmadany et al., 2023) introduced a benchmark covering 60 Arabic Natural Language Understanding (NLU) datasets across seven tasks, including WSD. Using the dataset by El-Razzaz et al. (2021), they reported a top F1-score of 76.68% with AraBERTv2, highlighting its effectiveness in MSA-based disambiguation.

Moreover, GPTAraEval (Khondaker et al., 2023) extended the evaluation to dialectal Arabic, revealing significant performance gaps between Modern Standard Arabic (MSA) and dialectal varieties when assessed using ChatGPT (GPT-3.5) and GPT-4. For WSD, they also utilized the dataset by El-Razzaz et al. (2021), in which ChatGPT achieved a best F1-score of 53.49% in a three-shot setting, reflecting the limitations of general-purpose LLMs in fine-grained disambiguation.

More recently, AraReasoner (Hasanaath et al., 2025) conducted a broad evaluation of reasoning-oriented LLMs, including DeepSeek models, across fifteen Arabic NLP tasks using various prompting and fine-tuning strategies. On the same dataset, their fine-tuned DeepSeek-R1-Q 14B model achieved up to 86.27% F1 score, demonstrating the effectiveness of task-specific adaptation.

In parallel, the ArabicNLP 2024 shared task (Khalilia et al., 2024) evaluated WSD systems on the SALMA dataset (Jarrar et al., 2023). The baseline model, a Target Sense Verification (TSV) system with a context window of 11 words, achieved the highest accuracy of 84.2%. Among the participants, Upaya obtained a top result of 77.82% using LLaMA-3-70B-Instruct with structural prompting. The shared task also evaluated Location Mention Disambiguation (LMD) using the IDRISI-DA dataset (Suwaileh et al., 2023a,b), which was created in two phases—first extracting location mentions, then disambiguating them. In this task, systems retrieved and reranked candidate toponyms from OpenStreetMap, with the best model achieving MRR@1 of 0.95.

Furthermore, EnhancedBERT (Kaddoura and Nassar, 2024b) introduces an ensemble BERT approach for Arabic WSD offering complementary benchmark.

Several other studies have also explored the performance of LLMs on Arabic NLP. However, most of these focus on specific applications or broader task suites that exclude WSD. In some cases, researchers develop their own datasets and conduct evaluations within that scope. Still, these efforts often lack generalization to fine-grained sense disambiguation, leaving essential gaps in systematic evaluation.

3 Arabic WSD Datasets: A Review

This section reviews key datasets for Arabic WSD, referred to here as Dataset A to Dataset F, with a focus on their construction methods and annotation schemes, summarized in Table 1.

Dataset A. Proposed by El-Razzaz et al. (2021), this dataset addresses the shortage of Arabic gloss-based resources by providing a public benchmark consisting of 15,549 senses for 5,347 unique Arabic words, extracted from the Modern Standard Arabic Dictionary. It frames Arabic WSD as a binary classification task, distinguishing between correct and incorrect glosses for a given word-in-context.

Dataset B. Proposed by Jarrar et al. (2023), SALMA is a novel Arabic sense-annotated cor-

Aspect	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E	Dataset F
Corpus size	15.5K tokens	34K tokens	3.7K sentence	27.5K sentence	28K pairs	167K pairs
Coverage	Single lemmas	Single lemmas	100 polysemous words	Single lemmas	Single lemmas	Single lemmas
Annotation	Gloss binary	Relatedness scores	Sense labeling	Gloss-based	Gloss binary	Gloss true/false
Construction method	Semi-Automatic	Manual	Manual, GPT-3.5	Fully Manual	Semi-Automatic	Semi-Automatic
Data type/ Domain	Dictionary examples	News and Media	Multi-domain	Multi-domain	Dictionary examples	Arabic Ont., Lex.
Source	MSA dictionary	Modern, Ghani	Web, GPT-3.5	DHDA dictionary	CAD dictionary	Arabic Ont., Lex.

Table 1: Summary of major Arabic WSD datasets (A–F). Abbrev.: Ont. = Ontology; Lex. = lexicography

pus containing around 34K tokens (approximately 29K annotated words), annotated simultaneously using two lexicons (Modern and Ghani). Unlike traditional binary methods, SALMA introduces a graded scoring system that assigns semantic relatedness scores to each sense (ranging from 1% to 100%). It also includes additional annotations for named entities.

Dataset C. Introduced by Kaddoura and Nassar (2024a), this dataset contains 3,670 context sentences representing 367 distinct senses across 100 carefully selected Arabic polysemous words. Sentences were manually collected from diverse online sources (e.g., news, medicine, finance) and supplemented with GPT-3.5-generated examples to cover less frequent senses.

Dataset D. Introduced by Saidi et al. (2023), WSDTN is a large-scale, manually annotated corpus of 27,530 Arabic sentences. It offers extensive semantic coverage. The annotation is based on the Doha Historical Dictionary of Arabic (DHDA).

Dataset E. Proposed at the KSAA-CAD shared task (Alshammari et al., 2024). This dataset provides approximately 28K Arabic gloss-context pairs sourced from the Contemporary Arabic Language Dictionary (CAD).

Dataset F. Al-Hajj and Jarrar (2021) introduced a significantly large dataset comprising approximately 167K context-gloss pairs extracted from the Arabic Ontology and the Birzeit lexicographic databases. The dataset is structured as a binary classification task (true/false).

4 Experimental Setup

4.1 Dataset Preparation

Dataset A [2] and Dataset B [3] were selected for evaluation as they are publicly available and offer complementary properties in terms of size, annotation schemes, and sense granularity.

Formatting and Preprocessing. Both datasets were organized into a consistent format comprising: (i) context sentences including the target word and candidate senses, (ii) ground-truth sense labels, and (iii) a dictionary mapping sense IDs to glosses. For Dataset B, tokens with invalid POS tags or missing semantic annotations were filtered to ensure cleaner input for disambiguation, and the sense with the highest score was treated as the correct label. The formatting strategy followed an approach similar to that used in the ArabicNLP 2024 shared task (Khalilia et al., 2024). Dataset examples are available in Appendix A.

Train-Test Splits. As shown in Table 2, custom 64/16/20 partitions were constructed for both datasets. For Dataset A, which contains a single target token per sentence, a random sentence-level split was applied to create training, development, and test sets. For Dataset B, where sentences may contain multiple targets, stratification was performed at the token level to ensure that 80% of annotated tokens were allocated to training and development, and 20% to testing. We ensured that no sentence appeared in both splits, preventing data leakage.

[2] https://github.com/MElrazzaz/Arabic-word-sense-disambiguation-bench-mark

[3] https://sina.birzeit.edu/salma/

Dataset	Train	Dev	Test	Total
Dataset A	9952	2487	3,110	15,549
Dataset B	18427	4691	5,781	28,899

Table 2: Token-level split statistics for the selected WSD datasets. Since no official splits are provided, custom partitions were created.

4.2 Model Selection

Two categories of models were compared in the evaluation:

- **Open-source LLMs:** LLaMA 3.1-8B (Grattafiori et al., 2024), Qwen 2.5-7B (Qwen et al., 2025), and Gemma 2-9B (Team et al., 2024).
- **Proprietary LLM:** GPT-4o (OpenAI et al., 2024).

The open-source models were evaluated under both zero-shot prompting and supervised fine-tuning. GPT-4o is used exclusively in the zero-shot setting, as fine-tuning this model is currently not feasible. This setup enables us to assess the effectiveness of instruction-tuned models for Arabic WSD and to examine how well relatively compact LLMs (7B–9B) perform compared to larger proprietary systems.

Zero-Shot Prompt:
You are ....
Sentence: ...
Target Word: ...
Possible Senses:
- Sense ID:, Definition:
- Sense ID:, Definition:
Correct Sense ID is:

Fine-Tuning TE:
{"**instruction**": "You are ...",
"**input**": "Sentence: '...' Target Word: '...' Possible Senses: [...]",
"**output**": "correct sense ID"}

Figure 1: Example formats for zero-shot prompting (left) and fine-tuning training examples (right).

4.3 Prompting and Fine-Tuning Strategies

Zero-Shot Prompting. All models were evaluated using a consistent prompt format that included the sentence, the target word, and a list of possible senses with their definitions. The models were instructed to select the correct sense ID (see Figure 1, left). Inference was performed using a mix of local deployments and API access, and we did not enforce deterministic decoding or temperature constraints. This stage aimed to assess how effectively large generative models can disambiguate senses in Arabic purely through instruction-following.

Supervised Fine-Tuning. For supervised adaptation, the open-source models were locally fine-tuned on the training splits of the benchmarks using parameter-efficient strategies, and specifically applied LoRA (Hu et al., 2021) in all experiments to reduce the memory footprint and training time. Training examples were formatted as instruction-style JSON objects containing the sentence, the target word, the candidate senses, and the correct label (see Figure 1, right). To handle long examples, truncation was applied non-uniformly: Dataset A samples were short, while for Dataset B sequences were retained up to 4,096 tokens, reducing training samples to 18,357. Training was performed on an NVIDIA L4 GPU with models loaded in 4-bit precision. Hyperparameters were tuned empirically to balance convergence speed and overfitting risk:

- **Dataset A**: epochs = 3, batch = 1 ($8\times$ accumulation), lr = 2×10^{-4}, max_len = 1024, LoRA rank = 32, $\alpha = 32$, dropout = 0.05, packing = `True`, eval_steps = 100
- **Dataset B**: epochs = 1, batch = 1 ($8\times$ accumulation), lr = 2×10^{-4}, max_len = 4096, LoRA rank = 16, $\alpha = 16$, dropout = 0.0, packing = `False`, eval_steps = 500
- **Common**: optimizer = `AdamW_8bit`, weight decay = 0.01, scheduler = `linear`, warmup steps = 50, gradient checkpointing = `True`, mixed precision = fp16/bf16 (auto), seed = 3407, load_in_4bit = `True`

4.4 Evaluation Metrics

Performance was evaluated using two complementary metrics: accuracy and macro-F1. Together, these measures provide a balanced perspective on how effectively the models addressed the task.

5 Results

Table 3 reports the results achieved by each model across both datasets.

5.1 Zero-Shot Results

In the zero-shot evaluation, GPT-4o achieved the highest performance, with similar results across both datasets. Among the open-source models, Gemma 2-9B performed best, particularly on Dataset B, where it surpassed the other models by a clear margin. Qwen 2.5-7B consistently outperformed LLaMA 3.1-8B, which had the lowest performance among the evaluated models.

Model	Dataset A				Dataset B			
	Zero-shot		Finetuning		Zero-shot		Finetuning	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Gemma 2-9B	65.34	50.64	89.39	81.72	72.46	56.45	87.23	67.80
LLaMA 3.1-8B	48.59	38.28	90.42	83.20	54.78	39.98	**88.51**	**69.41**
Qwen 2.5-7B	67.40	53.02	**90.77**	**83.98**	55.99	47.97	82.22	63.07
GPT-4o	79.16	67.92	–	–	79.55	64.23	–	–

Table 3: Accuracy and Macro-F1 scores of different models on Dataset A and Dataset B for Arabic WSD.

5.2 Fine-Tuning Results

Parameter-efficient fine-tuning led to substantial improvements across all open models. Qwen 2.5-7B achieved the best performance on Dataset A, while LLaMA-3.1 -8B, despite its lower zero-shot results, improved markedly with supervised adaptation and reached the highest scores on Dataset B. Gemma 2-9B also demonstrated significant gains across both datasets.

5.3 Results Analysis

We summarize four recurring phenomena observed in both datasets; for concreteness, we illustrate the patterns with LLaMA (see Figure 2):

- **In-set "close" vs. "distant" errors.** Close errors arise when glosses are near-paraphrases, whereas distant errors reflect semantic divergence. For الإفريقي, the gold "مَنْسوبٌ إلى إفْرِيقِيا" was predicted as "اسم منسوب إلى إفريقيا" a trivial close miss. In contrast, for هرب in "هرب من الحفلة بالنوم", the gold "هرب من مسئولياته: تنصل منها، تملص منها" was predicted as "هرب فلان في الأرض أبعد فيها", a distant error. Zero-shot runs were dominated by distant errors, as the datasets contain relatively few close glosses, (e.g., Gemma has on Dataset B: 1,465 distant vs. 5 close). However, fine-tuning markedly reduced them (e.g., LLaMA on Dataset A: 938 distant errors reduced to 291).

- **Invalid outputs (refusals + hallucinations).** Zero-shot models sometimes refused or produced non-existent IDs; LLaMA had 638 refusals on Dataset A and 539 on Dataset B. After fine-tuning, invalid outputs disappeared. Qwen also dropped from 1,277 refusals on Dataset B to 261 after tuning.

- **Effect of sense inventory size and dataset style.** Accuracy falls as the candidate set grows. Dataset A (dictionary-style; mostly 1–2 senses per token) is easier, whereas Dataset B (corpus-based; many items with 5+ senses) is harder.

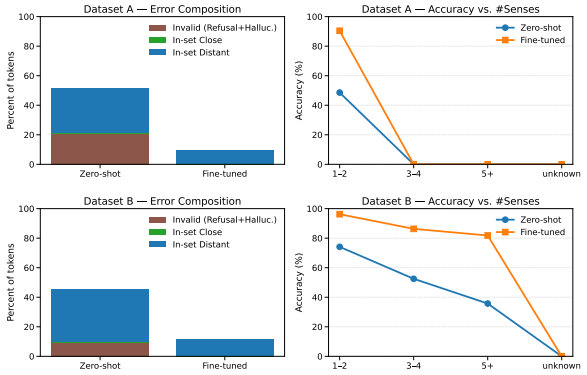

Figure 2: LLaMA-3.1-8B: zero-shot vs. fine-tuned on Dataset A and Dataset B. **Left:** 100% error composition (Invalid = refusal+hallucination; In-set Close; In-set Distant). **Right:** accuracy by number of candidate senses.

- **Difficult tokens.** In zero-shot, Dataset B concentrated errors on proper nouns and abstract terms (e.g., الجزائر, المركزية), while Dataset A's hardest cases were highly polysemous dictionary items (e.g., أمل, لبق). Fine-tuning removed zero-accuracy tokens in Dataset A, but some Dataset B tokens remained challenging (e.g., غرب).

6 Conclusion

This study benchmarked generative LLMs for Arabic WSD in zero-shot and fine-tuned settings across two public datasets. While GPT-4o led in zero-shot, parameter-efficient fine-tuning of open models consistently closed the gap and surpassed that baseline, yielding strong, reproducible results. Our analysis shows that factors such as sense-inventory size and error type drive performance differences and largely explain the gains from fine-tuning. Future work can expand to dialects.

Limitations

- **Dataset Scope.** This study focuses on two publicly available Modern Standard Arabic (MSA) datasets. The findings may not generalize to dialectal Arabic or other domains with different sense distributions and annotation practices.
- **Model Coverage.** We limited our evaluation to widely used multilingual LLMs. Arabic-centric models such as Jais and ALLaM, which may yield stronger performance, were not included due to stability and resource considerations.
- **Prompting Design.** To establish a clean zero-shot baseline, we used a minimal instruction-following prompt without examples or chain-of-thought reasoning. Richer prompting strategies (e.g., few-shot, reasoning heuristics, alternative gloss formats) could improve results but were left for future work.

Acknowledgments

We thank the Department of Electrical and Electronic Engineering, University of Khartoum, for their support.

References

Miuru Abeysiriwardana and Deshan Sumanathilaka. 2024. A survey on lexical ambiguity detection and word sense disambiguation. *Preprint*, arXiv:2403.16129.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. ArabGloss-BERT: Fine-tuning BERT on context-gloss pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 35–43, Held Online. INCOMA Ltd.

Sawsan Alqahtani, Hanan Aldarmaki, and Mona Diab. 2019. Homograph disambiguation through selective diacritic restoration. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 49–59, Florence, Italy. Association for Computational Linguistics.

Waad Alshammari, Amal Almazrua, Asma Al Wazrah, Rawan Almatham, Muneera Alhoshan, and Abdulrahman Alosaimy. 2024. KSAA-CAD shared task: Contemporary Arabic dictionary for reverse dictionary and word sense disambiguation. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 677–685, Bangkok, Thailand. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

M Soha Eid, Almoataz B Al-Said, Nayer M Wanas, Mohsen A Rashwan, and Nadia H Hegazy. 2010. Comparative study of rocchio classifier applied to supervised wsd using arabic lexical samples. In *Proceedings of the tenth conference of language engeneering (SEOLEC'2010), Cairo, Egypt*.

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A. Maghraby. 2021. Arabic gloss wsd using bert. *Applied Sciences*, 11(6).

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ahmed Hasanaath, Aisha Alansari, Ahmed Ashraf, Chafik Salmane, Hamzah Luqman, and Saad Ezzini.

2025. Arareasoner: Evaluating reasoning-based llms for arabic nlp. *Preprint*, arXiv:2506.08768.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023. SALMA: Arabic sense-annotated corpus and WSD benchmarks. In *Proceedings of ArabicNLP 2023*, pages 359–369, Singapore (Hybrid). Association for Computational Linguistics.

Sanaa Kaddoura and Reem Nassar. 2024a. A comprehensive dataset for arabic word sense disambiguation. *Data in Brief*, 55:110591.

Sanaa Kaddoura and Reem Nassar. 2024b. Enhanced-bert: A feature-rich ensemble model for arabic word sense disambiguation with statistical analysis and optimized data collection. *Journal of King Saud University - Computer and Information Sciences*, 36(1):101911.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The first Arabic natural language understanding shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 361–371, Bangkok, Thailand. Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *Preprint*, arXiv:2305.14976.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Rakia Saidi, Fethi Jarray, Asma Akacha, and Wissem Aribi. 2023. Wsdtn a novel dataset for arabic word sense disambiguation. In *Advances in Computational Collective Intelligence*, pages 203–212, Cham. Springer Nature Switzerland.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023a. IDRISI-D: Arabic and English datasets and benchmarks for location mention disambiguation over disaster microblogs. In *Proceedings of ArabicNLP 2023*, pages 158–169, Singapore (Hybrid). Association for Computational Linguistics.

Reem Suwaileh, Muhammad Imran, and Tamer Elsayed. 2023b. IDRISI-RA: The first Arabic location mention recognition dataset of disaster tweets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16298–16317, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Appendix

A Dataset Examples

Figure 3 and Figure 4 show illustrative examples from Dataset A and Dataset B, respectively, including the test-set sentence, the dictionary sense mapping, and the gold label.

Figure 3: Dataset A example.

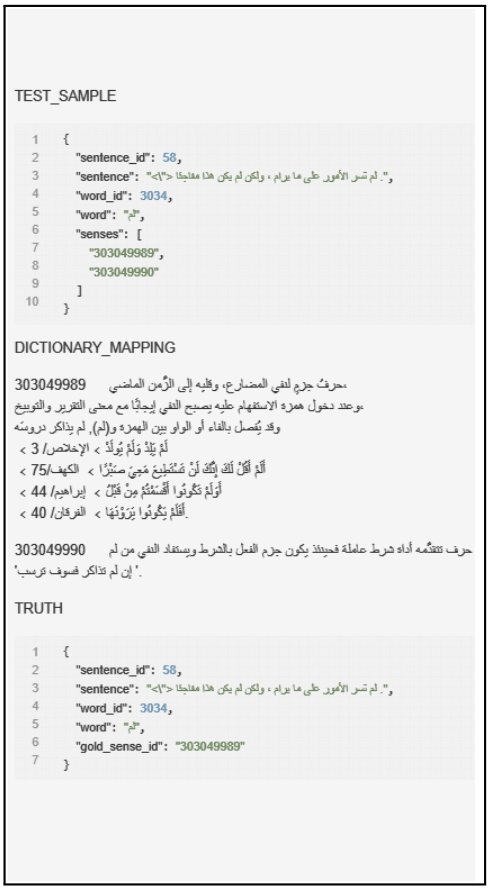

Figure 4: Dataset B example.