En-ViMedNER: An English-Vietnamese Parallel Biomedical Corpus with UMLS Semantic Type Annotations

Anonymous ACL submission

Abstract

Biomedical Named Entity Recognition (NER) is fundamental to healthcare AI applications, including clinical decision support and medical information extraction. While UMLS-annotated corpora such as MedMentions have driven progress in English biomedical NER, no comparable resource exists for Vietnamese. This paper presents En-ViMedNER, the first English-Vietnamese parallel biomedical NER corpus annotated with UMLS semantic types. The corpus contains 4,392 PubMed abstract pairs, 44,892 English-Vietnamese sentence pairs, and 202,949 aligned entity-mention pairs across 21 semantic types adapted from the MedMentions ST21pv dataset. To balance quality and scalability, we have constructed the corpus through automatic translation, expert post-editing, LLM-assisted label projection, and human verification and adjudication. We evaluate En-ViMedNER in two settings: (i) Vietnamese-input/Vietnamese-output biomedical NER and (ii) English-input/Vietnamese-output cross-lingual NER. For Vietnamese NER, we benchmark Vietnamese-supervised encoder models, English-supervised multilingual encoder models, and prompt-based LLMs. The best model achieves an F1 score of 52.70 on the test set and 53.78 on the mini-test set. For cross-lingual NER, we benchmark encoder-decoder models and prompt-based LLMs. The best model achieves an F1 score of 45.44 on the mini-test set. We publicly release our corpus, corpus construction pipeline, and baseline models to facilitate future Vietnamese biomedical NLP research. Given the semi-automated approach and the positive use cases of the corpus, we argue similar efforts should be made for some other low-resource languages.

1 Introduction

Biomedical named entity recognition (NER) is a core task in biomedical information extraction (IE) (Perera et al., 2020). It identifies mentions of biomedical concepts in unstructured text and supports downstream applications such as biomedical knowledge extraction, information retrieval, and question answering. While English biomedical NLP has benefited from large annotated resources and language models trained on biomedical text, Vietnamese biomedical NLP remains under-resourced, especially for entity recognition grounded in biomedical ontologies (Lee et al., 2019; Phan et al., 2022).

Previous work has shown that biomedical NER depends on both annotated corpora and biomedical knowledge resources. UMLS is a widely used biomedical ontology that provides semantic types for biomedical concepts (Bodenreider, 2004). MedMentions extended this line of work by linking PubMed abstracts to UMLS concepts and semantic types at scale (Mohan and Li, 2019). In parallel, BioBERT showed that pretraining on biomedical text can improve biomedical NER and related text mining tasks (Lee et al., 2019).

In Vietnamese, recent work has started to expand resources for biomedical and healthcare NLP, including PhoNER_COVID19 (Truong et al., 2021), ViHealthBERT (Minh et al., 2022), VietBioNER (Phan et al., 2022), ViMedNER (Duong et al., 2024), and VietMedNER (Le-Duc et al., 2025). However, these resources are still limited in scale, number of entity types, and are designed for spoken data or task-specific label sets, rather than a broad ontology such as UMLS.

This leaves two related gaps. First, to the best of our knowledge, there is still no publicly described English-Vietnamese biomedical NER resource that preserves UMLS semantic annotations, which are provided in English, for Vietnamese text. As a result, current Vietnamese biomedical NER research cannot benefit from cross-lingual transfer from English biomedical resources and interoperability with biomedical knowledge systems. Second, because biomedical documents and estab-

lished biomedical NER datasets/corpora are much more abundant in English than in Vietnamese, it is important to examine whether English biomedical supervision can support Vietnamese biomedical NER and English-to-Vietnamese cross-lingual NER, where a model converts the input English sentence into a Vietnamese sentence with NER tags. This setting is challenging because models must not only translate biomedical text, but also preserve entity boundaries and semantic labels across languages.

In this paper, we present **En-ViMedNER**[1], the first English-Vietnamese parallel biomedical NER corpus with UMLS-based semantic type annotations for Vietnamese text. Unlike existing Vietnamese biomedical NER resources, En-ViMedNER projects and preserves the semantic type annotations from MM-ST21pv[2] across aligned English-Vietnamese sentence pairs. Built from MedMentions through automatic translation, expert post-editing, LLM-assisted label projection, and manual verification and adjudication, the final corpus contains 4,392 PubMed abstract pairs, 44,892 English-Vietnamese sentence pairs, and 202,949 aligned entity-mention pairs. En-ViMedNER supports both Vietnamese biomedical NER and English-input/Vietnamese-output cross-lingual NER.

Our main contributions are as follows:

- We release En-ViMedNER, a parallel English-Vietnamese biomedical NER corpus with UMLS-based Vietnamese annotations.
- We provide comprehensive benchmarks for Vietnamese biomedical NER, comparing Vietnamese-supervised encoder fine-tuning, English-supervised multilingual encoder fine-tuning, and prompt-based LLMs.
- We evaluate English-input/Vietnamese-output cross-lingual biomedical NER using fine-tuned encoder-decoder models and prompt-based LLMs, and compare direct end-to-end generation with cascaded translation-and-NER pipelines.
- We introduce an LLM-as-a-judge evaluation protocol for cross-lingual NER generation, enabling more robust assessment of entity correctness under cross-lingual semantic variation.
- We publicly release the corpus, corpus construction pipeline, and benchmark setup to support reproducible research in Vietnamese biomedical NLP.

Although this work focuses on Vietnamese, the pipeline can be adapted to construct UMLS-based biomedical NER corpora for other low-resource languages, provided that reliable biomedical MT, target-language medical review, and span validation are available.

2 Background and Related Work

In this section, we introduce the key concepts and resources underlying our work.

2.1 UMLS and Semantic Types

The Unified Medical Language System (UMLS), maintained by the U.S. National Library of Medicine (D.A.B.Lindberg et al., 1993), is a comprehensive biomedical knowledge resource that integrates over 200 source vocabularies[3] to facilitate interoperability across biomedical information systems. One of the three UMLS knowledge sources is the Semantic Network. The Semantic Network is composed of 127 semantic types as nodes and 54 semantic relations as links between those nodes[4]. Semantic types enable standardized classification of all biomedical concepts in the UMLS Metathesaurus, making them well-suited as label categories for biomedical NER.

2.2 MedMentions Corpus

MedMentions (Mohan and Li, 2019) is a large-scale biomedical corpus designed to promote research in biomedical named entity recognition and concept normalization. The corpus consists of 4,392 PubMed abstracts randomly selected from papers published in 2016. Each abstract was manually annotated by professional experts who identified entity mentions and linked them to UMLS concepts using Concept Unique Identifiers (CUIs). In total, the corpus contains 352,496 mentions corresponding to 34,724 unique UMLS concepts spanning 126 of 127 semantic types, with a reported inter-annotator agreement of 97.3%.

[1] Our En-ViMedNER dataset is publicly available at: anonymous url

[2] MM-ST21pv contains annotations of 21 semantic types (Mohan and Li, 2019), while the full MedMentions corpus contains annotations of 126 semantic types (Fraser et al., 2019).

[3] https://uts.nlm.nih.gov/uts/umls/home

[4] https://www.ncbi.nlm.nih.gov/books/NBK9679/

In addition to the full corpus, Mohan and Li released MedMentions ST21pv (MM-ST21pv), a filtered version retaining only mentions mapped to the 21 clinically relevant semantic types in the ST21pv subset across 18 source ontologies. This subset provides a more tractable label space while maintaining broad biomedical coverage. Our work derives from MM-ST21pv, translating the English abstracts into Vietnamese and projecting the semantic type annotations onto the translated text.

2.3 Vietnamese Biomedical NER

Research on Vietnamese biomedical NER remains limited compared to English. Ngo et al. (2019) constructed a Vietnamese medical terminology resource integrable into UMLS, mapping approximately 11,383 disease terms and 7,146 drug names to English corresponding to UMLS concepts. While valuable for lexical lookup, this resource does not provide annotated text for training NER models.

More recently, Duong et al. (2024) introduced a Vietnamese medical NER dataset named ViMedNER with initial benchmarks using pre-trained language models. Le-Duc et al. (2025) presented a greater effort with 18 entity types across 9,000 sentences of medical speech transcripts in a dataset named VietMedNER. Their experiments demonstrated that encoder-only architectures (BERT, XLM-R, DeBERTa) consistently outperform sequence-to-sequence models (BART, T5) for Vietnamese medical NER. Notably, VietMedNER employed a large test set relative to the training set to obtain more statistically significant evaluation results when leveraging large pre-trained models.

However, no existing Vietnamese biomedical NER corpus provides annotations with UMLS semantic types. Our work addresses this gap by constructing the first such resource derived from MedMentions ST21pv.

2.4 Cross-Lingual Transfer

Cross-lingual transfer is a practical strategy for biomedical NER because high-quality annotations are often concentrated in English, while target languages such as Vietnamese remain under-resourced (Miftahutdinov et al., 2020). Prior studies show that multilingual encoders such as mBERT and XLM-R can support cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020a), including in biomedical and clinical NER (Hakala and Pyysalo, 2019; Miftahutdinov et al., 2020).

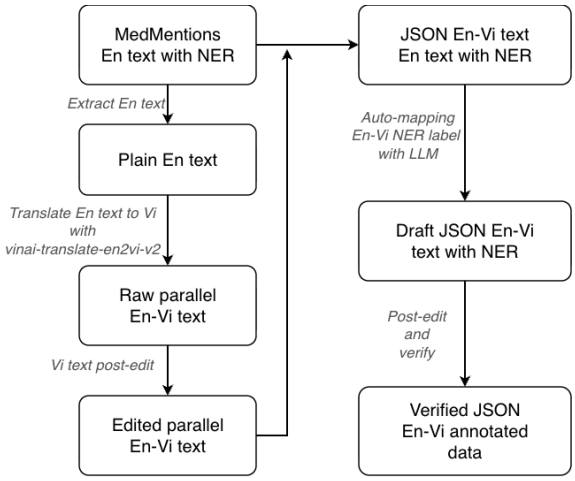

Figure 1: Pipeline for constructing English-Vietnamese biomedical NER data from MedMentions.

However, transfer performance can be affected by language distance, domain mismatch, terminology variation, and entity-boundary differences (Rivera-Zavala and Martínez, 2021; Lancheros et al., 2025). These issues are especially relevant for English-Vietnamese biomedical NER, where biomedical terms and span boundaries may change across languages.

Translation-based methods provide another way to reduce annotation cost. Prior work has projected entity labels through machine translation or aligned spans in translated clinical data (Jain et al., 2019; Schäfer et al., 2022). Other studies compare direct transfer with translation-based NER, showing that translation can be useful but requires careful preservation of entity spans and labels (Gaschi et al., 2023). More recently, cross-lingual NER has also been formulated as a generation task, where models produce the target-language sentence together with explicit entity tags (Yang et al., 2022).

Building on these directions, we construct En-ViMedNER as a parallel English-Vietnamese biomedical NER corpus and evaluate it in two complementary settings. First, we examine whether English biomedical supervision can transfer to Vietnamese NER through multilingual encoder fine-tuning. Second, we study cross-lingual NER, comparing prompt-based LLM generation with fine-tuned encoder-decoder models under both direct and cascaded translation-and-NER pipelines.

3 Corpus Construction

We construct **En-ViMedNER** from MM-ST21pv using three main steps: translation and post-editing,

Split	Abstract pairs	Sentence pairs	Aligned entity-mention pairs
train	2,635	27,013	122,017
dev	878	8,932	40,844
test	879	8,947	40,088
mini-test*	–	450	2,085
Total	4,392	44,892	202,949

Table 1: Statistics of the En-ViMedNER dataset across train, development, and test splits. **Note**: The mini-test set is randomly sampled from the test set for human evaluation and is not counted as an additional split.

label projection, and human verification, as shown in Figure 1.

3.1 Machine Translation and Medical Expert Post-editing

We extract English sentences from MM-ST21pv and remove overlapping or unlabeled sentences, resulting in 44,892 sentences. These sentences are translated into Vietnamese using vinai-translate-en2vi-v2 (Nguyen et al., 2022). All Vietnamese translations are then post-edited by domain experts[5] to improve medical terminology and sentence quality. This step produces a fully post-edited English-Vietnamese parallel corpus.

3.2 LLM-assisted Label Projection

We then convert the parallel sentences into JSON templates compatible with Label Studio. Each template keeps the original English entity spans and labels, while the Vietnamese annotation fields are left blank for label projection (Appendix A). We use gpt-5.2[6] (OpenAI, 2026a) to map each English entity to its corresponding Vietnamese span and assign the same semantic label.

After projection, we apply automatic checks by comparing English source entities with projected Vietnamese annotations to detect potential mislabels, primarily due to untranslated entities or omitted abbreviations. The flagged cases are manually reviewed and corrected. At the corpus scale, these automatic checks identify about 1.77% of sentence pairs that contain Vietnamese mislabel entities.

We convert the verified entity spans into BIO-tagged data for the experiments. Following the original MedMentions split, we divide the corpus into **train**, **dev**, and **test** sets with an approximate 60/20/20 ratio, and sample 450 test sentences as a **mini-test** set for human evaluation. Table 1 reports the final corpus statistics.

3.3 Human Verification and Adjudication

We assess the quality of the automatically projected Vietnamese annotations on the mini-test set (2,085 entity pairs). Before manual correction, two authors independently reviewed the projected Vietnamese spans and labels against the English source annotations and the post-edited Vietnamese translations.

They checked whether each projected span matched the intended biomedical entity and whether its character boundaries were correct. Detailed evaluation criteria are provided in Appendix B.1.

Inter-annotator agreement before discussion was measured using Cohen's κ (Cohen, 1960) at the sentence level. For each of the 450 sentences, annotators judged whether an error occurred under each projection criterion. The overall pooled κ is 0.83, computed over sentence-level judgments, indicating substantial to almost-perfect agreement according to the Landis and Koch scale (Landis and Koch, 1977). Detailed scores are reported in Appendix B.2.

After discussion, all disagreements and suspected error cases were resolved by consensus, and confirmed errors were manually corrected for the final version (Appendix B.3). Based on this consensus audit, the GPT-projected annotations reached 98.32% exact-match accuracy before correction. Thus, at the mini-test set scale, manual correction improved the audited annotations by about 1.68 percentage points.

4 Experimental

4.1 Experimental Settings

We organize our experiments into two main tasks:

Task 1 - Vietnamese NER. This task tests whether En-ViMedNER provides useful Vietnamese supervision beyond English-supervised transfer and prompt-based LLMs. We compare fine-tuned encoder models (Track 1.1) and prompt-based LLMs (Track 1.2).

Track 1.1 - Fine-tuned encoder models: We evaluate fine-tuned encoder models under two settings. First, in the Vietnamese-supervised setting,

[5] Final-year Vietnamese medical students or graduates with Doctor of Medicine degrees.

[6] With none reasoning effort, medium verbosity, and top-p 0.98. We choose this model because, in a preliminary test on one abstract (98 labels), it achieves 100% exact match accuracy, compared with 91.8% for gemini-3.0-pro-preview (Google, 2026a).

PhoBERT, XLM-R, ViHealthBERT, and ViPubMedDeBERTa are fine-tuned on the Vietnamese training split and evaluated on the Vietnamese test split. Second, in the English-supervised setting, multilingual XLM-R models are fine-tuned on the English training split and evaluated directly on the same Vietnamese test split. Results are reported on the full test set, with mini-test scores included for comparison with prompt-based LLMs in Track 1.2.

Track 1.2 - Prompt-based LLMs: We evaluate prompt-based LLMs on the mini-test set. We include both open-source LLMs, such as Qwen and SEA-LION, and closed-source frontier LLMs, such as GPT and Gemini. In this setting, each model receives a Vietnamese sentence and is asked to reproduce the sentence with BIO biomedical NER tags. We evaluate zero-shot, 1-shot, and 3-shot prompting, where few-shot examples are retrieved from the Vietnamese labeled training pool using embedding similarity.

Overall, these experiments test whether supervised fine-tuning on En-ViMedNER is beneficial for Vietnamese biomedical NER, compared with prompt-based LLMs.

Task 2 - Cross-lingual NER: English-input/Vietnamese-output. This task evaluates English-to-Vietnamese cross-lingual NER, where models generate a Vietnamese translation with biomedical entity tags from a plain English sentence. We compare fine-tuned encoder-decoder models (Track 2.1) and prompt-based LLMs (Track 2.2) on the mini-test set.

Track 2.1 - Fine-tuned multilingual encoder-decoder models: We fine-tune umT5 and NLLB-200-distilled-600M on the parallel data with inline NER tags. We compare direct end-to-end cross-lingual NER with cascaded pipelines that separate translation and NER into sequential steps. We define three settings:

- **M1: Direct Trans.+NER**: direct end-to-end fine-tuning, where the input is a plain English sentence and the output is a tagged Vietnamese sentence.
- **M2: Trans.→NER**: a cascaded pipeline where the **Trans. only** component first translates the plain English input into a plain Vietnamese sentence, and the **NER only** component then tags the translated Vietnamese sentence.
- **M3: NER→Trans.**: a cascaded pipeline where the **NER only** component first tags the plain English input, and **Tag-aware Trans.** component then translates the tagged English sentence into a tagged Vietnamese sentence.

An overview diagram (Figure 5) and details of **Trans. only/NER only/Tag-aware Trans.** components are provided in Appendix C.

Track 2.2 - Prompt-based LLMs: We evaluate prompt-based LLMs under the same English-input/Vietnamese-output setting. We evaluate both open-source and closed-source LLMs under zero-shot, 1-shot, and 3-shot prompting, where few-shot examples are retrieved from the parallel labeled training pool using embedding similarity.

Overall, these experiments test whether supervised fine-tuning on En-ViMedNER is beneficial for cross-lingual NER, compared with prompt-based LLMs.

4.2 Models

We evaluate six distinct groups of models (see Appendix D for details of the model setups, training configurations, hyperparameters, and LLM's prompt templates):

- **Monolingual encoders.** We use PhoBERT-base/base-v2/large, which are RoBERTa-style models pre-trained for Vietnamese (Nguyen and Tuan Nguyen, 2020).
- **Vietnamese biomedical and healthcare encoders.** We use ViHealthBERT-word/syllable, domain-specific pre-trained language models for Vietnamese healthcare text mining (Minh et al., 2022); ViPubMedDeBERTa-xsmall/base, which are DeBERTa-style models pre-trained for Vietnamese biomedical text (Tran-Tien et al., 2023).
- **Multilingual encoders.** We use XLM-R-base/large, multilingual masked language models pre-trained on 100 languages (Conneau et al., 2020b).
- **Multilingual encoder-decoder models.** We use umT5-base, a multilingual T5-style encoder-decoder model (Chung et al., 2023), and NLLB-200-distilled-600M, a distilled model from the NLLB-200 translation family (NLLB Team et al., 2022).
- **Open-source LLMs.** We use Qwen2.5-3B/7B-Instruct (Team, 2024), and Qwen-SEA-LION-v4-8B-VL, a Southeast Asia-oriented model adapted from Qwen (AI Singapore, 2025).
- **Closed-source LLMs.** We use gpt-5.4 and gemini-3.1-flash-lite-preview as frontier

closed-source LLMs through their official APIs (OpenAI, 2026b; Google, 2026b).

4.3 Evaluation Metrics

For Task 1, we report entity-level precision, recall, and F1 using seqeval (Nakayama, 2018). For Task 2, we evaluate both translation quality and biomedical NER performance. We use SacreBLEU (Post, 2018) to measure the generated Vietnamese sentence against the reference Vietnamese translation. For NER performance, we report two metrics: **Exact-F1** and **Judge-F1**. Exact-F1 is computed with seqeval after converting inline entity tags into BIO sequences, requiring exact surface-form and label matching. Judge-F1 is computed from the LLM-as-a-judge protocol described in the next subsection.

LLM-as-a-judge Evaluation. For the evaluation of Track 2.1 and Track 2.2, exact span matching is often too restrictive. In these settings, the input sentence is written in English, while the model output is a Vietnamese sentence with NER tags. As a result, the predicted entity span may be a valid translation or paraphrase of the gold Vietnamese entity, rather than an exact text match. This limits direct F1 computation against the gold Vietnamese annotation, especially when multiple but semantically equivalent translations are possible.

To address this issue, we introduce an LLM-as-a-judge evaluation protocol that applies to experiments where the input and output languages differ. We use gpt-5.4-mini with zero-shot prompting as the judge. Given a model prediction and its extracted entity list, together with the gold sentence and gold entity list, the judge is asked to match each gold entity with the corresponding predicted entity, if any. For each gold entity, the judge evaluates whether the predicted text span and entity label are correct. Each case is categorized as *correct*, *incorrect*, or *miss*. The judge is also asked to identify extra predicted entities that do not correspond to any gold entity (see judging prompt at Appendix H). We finally convert the judge decisions into entity-level F1 counts as follows:

- Exact span and tag match: counted as one TP.
- Any span or tag mismatch, including cases where both are incorrect, is counted as one FP for the prediction and one FN for the corresponding gold entity.
- Missed gold entity: counted as one FN.
- Extra predicted entity: counted as one FP.

To verify the reliability of this protocol, we measure human-LLM agreement on 100 randomly sampled outputs from the NLLB-M1 setting. Human annotators apply the same criteria as the LLM judge, covering entity-span matching, label matching, missed entities, and extra predictions, and we report Cohen's κ to assess the consistency of the LLM judge with human judgment.

5 Results

5.1 Task 1 - Vietnamese NER

Table 2 compares three settings for Vietnamese NER: Vietnamese-supervised encoder fine-tuning, English-supervised multilingual encoder fine-tuning, and prompt-based LLMs.

Vietnamese supervision provides a clear advantage. Vietnamese-supervised encoder models consistently outperform English-supervised multilingual encoder models on both evaluation sets. On the full test set, ViPubMedDeBERTa-base achieves 52.70 F1, outperforming the best English-supervised model, XLM-R-large, by 16.16 points. The same pattern holds on the mini-test set, where ViPubMedDeBERTa-xsmall obtains 53.78 F1 compared with 37.95 F1 for XLM-R-large. These results indicate that multilingual pretraining and English supervision cannot replace target-language annotations, highlighting the value of En-ViMedNER for Vietnamese biomedical entity boundaries, terminology, and labels.

Prompting improves with examples but remains below fine-tuning. Few-shot prompting improves both open-source and closed-source LLMs, but the gains remain insufficient to match supervised encoder fine-tuning. For example, gemini-3.1-flash-lite-preview improves from 12.70 F1 in the zero-shot setting to 36.47 F1 with 3-shot prompting. Among open-source LLMs, Qwen-SEA-LION-v4-8B-VL improves from 1.59 to 13.09 F1. Nevertheless, the best prompted model, gemini-3.1-flash-lite-preview-3shot, remains 17.31 F1 points behind the best Vietnamese-supervised encoder. These results suggest that in-context examples help LLMs follow the BIO tagging format, but do not replace task-specific Vietnamese biomedical supervision.

Overall, the findings from Task 1 demonstrate that En-ViMedNER constitutes a valuable resource for developing robust Vietnamese biomedical named entity recognition (NER) systems, rather

Model	mini-test set			Full test set		
	Precision	Recall	F1	Precision	Recall	F1
Vietnamese-supervised setting						
PhoBERT-base	54.28	51.83	53.03	52.65	49.98	51.28
PhoBERT-base-v2	53.72	51.68	52.68	52.60	50.42	51.49
PhoBERT-large	53.41	51.63	52.51	52.19	50.84	51.50
XLM-R-base	53.53	52.12	52.81	50.68	48.92	49.78
XLM-R-large	54.91	51.63	53.22	52.66	50.29	51.45
ViHealthBERT-word	53.70	51.25	52.45	52.72	50.51	51.59
ViHealthBERT-syllable	52.37	50.00	51.16	50.42	48.47	49.42
ViPubMedDeBERTa-xsmall	54.00	53.56	**53.78**	52.35	52.74	52.54
ViPubMedDeBERTa-base	54.03	53.22	53.62	52.86	52.54	**52.70**
English-supervised setting						
XLM-R-base	30.26	37.72	33.58	27.80	35.48	31.12
XLM-R-large	35.14	41.26	37.95	33.54	40.13	36.54
Open-source LLMs						
Qwen2.5-3B-Instruct-0shot	4.03	0.83	1.37			
Qwen2.5-3B-Instruct-1shot	8.08	2.09	3.32			
Qwen2.5-3B-Instruct-3shot	9.76	2.14	3.50			
Qwen2.5-7B-Instruct-0shot	2.14	0.24	0.44			
Qwen2.5-7B-Instruct-1shot	9.57	2.62	4.12			
Qwen2.5-7B-Instruct-3shot	12.25	2.72	4.45			
Qwen-SEA-LION-v4-8B-VL-0shot	8.61	0.87	1.59			
Qwen-SEA-LION-v4-8B-VL-1shot	16.10	6.65	9.41			
Qwen-SEA-LION-v4-8B-VL-3shot	21.22	9.47	13.09			
Closed-source LLMs						
gpt-5.4-0shot	11.79	12.99	12.36			
gpt-5.4-1shot	20.04	22.50	21.20			
gpt-5.4-3shot	32.19	36.26	34.11			
gemini-3.1-flash-lite-preview-0shot	13.16	12.26	12.70			
gemini-3.1-flash-lite-preview-1shot	23.55	24.92	24.21			
gemini-3.1-flash-lite-preview-3shot	35.10	37.95	36.47			

Table 2: Performance comparison of encoder fine-tuning and prompt-based LLMs on Vietnamese biomedical NER. The input is a plain-text Vietnamese sentence, and the output is Vietnamese NER annotations. Bold numbers indicate the highest F1 score within each evaluation set.

than serving solely as an evaluation benchmark.

5.2 Task 2 - Cross-lingual NER

Table 3 compares fine-tuned encoder-decoder models and prompt-based LLMs on the mini-test set. We report entity-level precision, recall, and Judge-F1 using the LLM-as-a-judge protocol, together with Exact-F1 based on exact entity matching against the gold annotations. The reliability of the judging protocol is further supported by human-LLM agreement, with Cohen's $\kappa = 0.87$.

Fine-tuned encoder-decoder models outperform prompting. Fine-tuning on En-ViMedNER parallel data gives the strongest overall performance. Under Judge-F1, the best fine-tuned setting is NLLB-M1, which reaches 45.44 F1 and outperforms the best closed-source LLMs, including gpt-5.4-3shot and gemini-3.1-flash-lite-preview-3shot, by 2.80 and 3.89 points, respectively. The same trend holds under the stricter Exact-F1 metric: NLLB-M3 achieves the highest Exact-F1 score of 15.98, above the best closed-source LLM result of 10.40. These results suggest that paired English inputs and Vietnamese tagged outputs provide useful supervision for learning both translation and biomedical entity transfer.

Neither end-to-end nor cascaded pipelines are uniformly superior. Under Judge-F1, direct end-to-end generation performs best for both models: umT5-M1 achieves 26.88 F1, while NLLB-M1 achieves 45.44 F1. However, under Exact-F1, the **NER→Trans.** pipeline performs best for both models, reaching 7.33 for umT5-M3 and 15.98 for NLLB-M3. This suggests that direct generation is more effective when semantically equivalent entity translations are accepted, whereas the tag-first cascaded pipeline may better preserve exact entity forms. Therefore, the relative advantage of end-to-end and cascaded pipelines depends on the evaluation criterion.

Open-source LLMs struggle with both tagging quality and output validity. Open-source LLMs remain substantially weaker than fine-tuned encoder-decoder models and closed-source LLMs under Judge-F1. The best open-source result is Qwen-SEA-LION-v4-8B-VL-3shot, which reaches 18.43 Judge-F1 and 9.85 Exact-F1, still far below the best fine-tuned encoder-decoder results (NLLB-M1). A major issue is invalid formatting: open-source LLMs often fail to produce valid Vietnamese NER-tagged outputs, with invalid counts ranging from 114 to 292 out of 450 examples. This

Setting/Model	Metrics					
	BLEU	Prec.	Rec.	Judge-F1	Exact-F1	Inv.
Fine-tuned multilingual encoder-decoder models						
umT5-M1: Direct Trans.+NER	29.85	24.41	29.90	26.88	6.01	0
umT5-M2: Trans.→NER	35.98	32.80	19.71	24.62	5.17	0
umT5-M3: NER→Trans.	31.15	30.64	23.56	26.64	7.33	0
NLLB-M1: Direct Trans.+NER	48.63	44.31	46.63	**45.44**	15.77	0
NLLB-M2: Trans.→NER	51.83	43.77	38.33	40.87	13.64	0
NLLB-M3: NER→Trans.	46.90	41.32	36.95	39.02	**15.98**	0
Open-source LLMs						
Qwen2.5-3B-Instruct-0shot	35.81	8.73	2.14	3.43	0.91	114
Qwen2.5-3B-Instruct-1shot	36.39	8.98	2.77	4.23	1.80	178
Qwen2.5-3B-Instruct-3shot	38.73	11.86	4.17	6.18	3.33	158
Qwen2.5-7B-Instruct-0shot	40.79	7.93	1.26	2.18	1.99	194
Qwen2.5-7B-Instruct-1shot	36.71	15.20	3.79	6.06	3.92	201
Qwen2.5-7B-Instruct-3shot	43.66	22.28	8.16	11.94	5.34	115
Qwen-SEA-LION-v4-8B-VL-0shot	50.06	21.00	3.88	6.55	6.17	292
Qwen-SEA-LION-v4-8B-VL-1shot	49.99	22.59	8.06	11.88	9.19	250
Qwen-SEA-LION-v4-8B-VL-3shot	52.89	30.23	13.25	18.43	9.85	228
Closed-source LLMs						
gpt-5.4-0shot	56.59	28.94	35.68	31.96	4.43	0
gpt-5.4-1shot	**56.60**	33.24	41.48	36.91	7.33	0
gpt-5.4-3shot	53.34	38.74	47.42	42.64	9.84	0
gemini-3.1-flash-lite-preview-0shot	55.03	30.03	34.14	31.95	4.43	0
gemini-3.1-flash-lite-preview-1shot	52.18	34.70	42.39	38.17	8.05	0
gemini-3.1-flash-lite-preview-3shot	48.80	38.59	45.00	41.55	10.40	0

Table 3: Performance comparison across encoder-decoder fine-tuning and prompt-based LLMs on the mini-test set. The input is a plain-text English sentence, and the output is Vietnamese NER annotations. **Inv.** denotes the number of generated sentences without valid entity tags. Bold numbers indicate the highest value in the corresponding metric column.

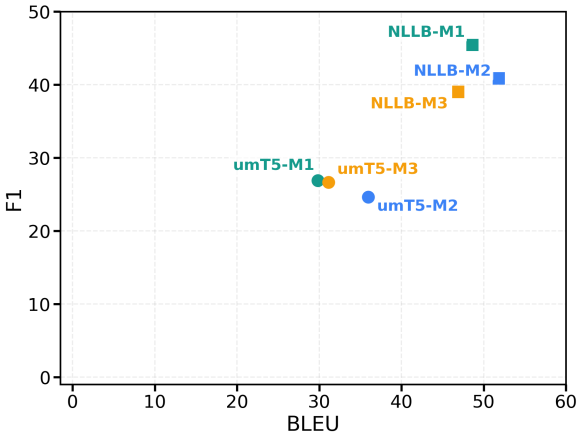

Figure 2: Relationship between translation quality and NER performance for encoder-decoder cross-lingual NER on the mini-test set. Circles indicate umT5-base, squares indicate NLLB-200-distilled-600M, and each point represents one experimental setting. *Note*: Colors are used only as a secondary visual cue to group the same setting across models; the figure remains interpretable through marker shapes and direct labels.

suggests that these models often fail to consistently follow the required biomedical tagging format in the cross-lingual generation setting.

Higher BLEU does not always imply better NER. Figure 2 shows that translation quality and NER performance are not perfectly aligned. For example, NLLB-M2 obtains the highest BLEU score among the main NLLB systems, but its F1 is lower than NLLB-M1. A similar pattern appears for umT5: M2 has the highest BLEU score, but M1 gives the best F1. This indicates that good sentence-level translation is not enough for biomedical NER generation. The model must also preserve entity boundaries and labels during generation.

Overall, the findings from Task 2 demonstrate that En-ViMedNER's parallel English-Vietnamese annotations provide a valuable foundation for cross-lingual biomedical NER, enabling models to jointly learn translation, entity-boundary preservation, and biomedical label transfer.

6 Conclusion and Future Work

We presented En-ViMedNER, the first parallel English-Vietnamese biomedical NER corpus with UMLS semantic annotations, comprising 4,392 PubMed abstracts and 202,949 aligned entity-mention pairs across 21 UMLS semantic types. The corpus is constructed via a hybrid pipeline of machine translation, expert post-editing, LLM-assisted projection, and human verification (Cohen's $\kappa = 0.83$, entity-level exact match = 98.32%).

Our experiments demonstrate the value of En-ViMedNER in two settings. For Vietnamese biomedical NER, Vietnamese-supervised encoder models consistently outperform both English-supervised multilingual encoders and prompt-based LLMs, showing that multilingual pretraining and prompting cannot replace target-language biomedical supervision. For cross-lingual NER, fine-tuned encoder-decoder models outperform prompt-based LLMs, indicating that the parallel annotations provide useful supervision for jointly learning translation, entity-boundary preservation, and biomedical label transfer. The results further show that translation quality alone is insufficient for cross-lingual biomedical NER, and that end-to-end and cascaded pipelines offer different advantages depending on the evaluation criterion.

Future work will explore extensions to other low-resource Southeast Asian languages, particularly Indonesian and Thai, where emerging medical NER resources provide useful starting points for evaluating the portability of our construction pipeline. We will also investigate additional experimental settings involving open-source LLM fine-tuning, LLM-as-a-judge reliability (Wang et al., 2026), and cascaded medical translation-NER pipelines.

Limitations

Despite our efforts to ensure corpus quality through expert post-editing of translations, as well as review and correction after LLM-assisted label projection, there are limitations that users of En-ViMedNER should be aware of.

- **Sentence-level vs. abstract-level context**: Some Vietnamese translations only make sense in a broad context of an abstract. For example, the translation *animals –> động vật thí nghiệm* assumes a laboratory-experiment context, which is available at the abstract level but not at the sentence level. Because our annotation and evaluation operate at the sentence level, the meaning of some words may appear less natural than they would in full context.
- **Surface-form ambiguity**: There are cases when the same surface form appears multiple times in a sentence, some are translated but some are kept as the original text. For example, *HCC* appears twice in sentence *Mep1A is overexpressed in most HCC and induces HCC cell migration and invasion.*, the first *HCC* is kept as-is while the second *HCC* is translated to *ung thư gan*. Cases like this may confuse trained NER models and reduce surface-level learnability.
- **Domain coverage**: En-ViMedNER is built from PubMed abstracts, which are formal biomedical literature. Therefore, models trained on En-ViMedNER may underperform on less formal text like clinical notes and patient-facing health communication.
- **Downstream use**: Models trained on En-ViMedNER should not be assumed to perform reliably on clinical notes, patient-facing text, or safety-critical medical applications without additional validation.

Ethical considerations

En-ViMedNER follows the licensing conditions of both MedMentions and UMLS. Since MedMentions is released under CC0, we can redistribute the translated dataset. We only retain UMLS semantic type codes (e.g., T047) as reference labels, and do not publish UMLS concept names, definitions, or hierarchy information. Users who need detailed content must access it through their own UMLS accounts under the official UMLS license.

The corpus is derived from PubMed abstracts and MedMentions rather than clinical notes or patient records, reducing the risk of personally identifying health information.

Use of AI tools

We used Grammarly[7] for grammar checking and readability improvement during manuscript preparation. All scientific claims, analyses, and final writing decisions remain the responsibility of the authors.

References

AI Singapore. 2025. Qwen-SEA-LION-v4-8B-VL. https://huggingface.co/aisingapore/Qwen-SEA-LION-v4-8B-VL. Hugging Face model card. Accessed: 2026-04-24.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. ICLR 2023 poster.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

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

D.A.B.Lindberg, B.L.Humphreys, and A.T.McCray. 1993. The Unified Medical Language System. *Yearbook of Medical Informatics*, pages 41–51.

[7] https://app.grammarly.com/

Pham Van Duong, Tien-Dat Trinh, Minh-Tien Nguyen, Huy-The Vu, Minh Chuan Pham, Tran Manh Tuan, and Le Hoang Son. 2024. ViMedNER: A medical named entity recognition dataset for Vietnamese. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, 11(4).

Kathleen C Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. 2019. Extracting UMLS Concepts from Medical Text Using General and Domain-Specific Deep Learning Models. In *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 157–167.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical NER: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada. Association for Computational Linguistics.

Google. 2026a. Gemini 3.0 Preview. https://ai.google.dev/gemini-api/docs/models. Accessed: 2026-05-13.

Google. 2026b. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed: 2026-04-24.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China. Association for Computational Linguistics.

B. S. Lancheros, G. Corpas Pastor, and R. Mitkov. 2025. Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain. *Language Resources & Evaluation*, 59(2):665–684.

J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74.

Khai Le-Duc, David Thulke, Hung-Phong Tran, Long Vo-Dang, Khai-Nguyen Nguyen, Truong-Son Hy, and Ralf Schlüter. 2025. Medical Spoken Named Entity Recognition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 724–783, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2020. On biomedical named entity recognition: Experiments in interlingual transfer for clinical and social media texts. In *Advances in Information Retrieval*, volume 12036 of *Lecture Notes in Computer Science*, pages 281–288. Springer.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. ViHealthBERT: Pre-trained language models for Vietnamese in health text mining. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 328–337, Marseille, France. European Language Resources Association.

Sunil Mohan and Donghui Li. 2019. MedMentions: A large biomedical corpus annotated with UMLS concepts. In *Proceedings of the 2019 Conference on Automated Knowledge Base Construction (AKBC 2019)*, Amherst, Massachusetts, USA.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval. Accessed: 2026-04-07.

The Quyen Ngo, My Linh Ha, Thi Minh Huyen Nguyen, Thi Mai Huong Hoang, and Viet Hung Nguyen. 2019. Towards a UMLS-integratable Vietnamese medical terminology. In *International Conference of the Pacific Association for Computational Linguistics*, pages 388–399, Singapore. Springer.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Nho Trung Nguyen, Nhat Quang Nguyen, and Van Huy Nguyen. 2025. Vietnamese_Embedding: Embedding model in Vietnamese language. Accessed: 2026-03-11.

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2026a. GPT-5.2 Model. https://developers.openai.com/api/docs/models/gpt-5.2. Accessed: 2026-04-24.

OpenAI. 2026b. GPT-5.4 Model. https://developers.openai.com/api/docs/models/gpt-5.4. Accessed: 2026-04-24.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, Volume 8 - 2020.

Uyen Phan, Phuong N.V Nguyen, and Nhung Nguyen. 2022. A named entity recognition corpus for Vietnamese biomedical texts to support tuberculosis treatment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3601–3609, Marseille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

R. M. Rivera-Zavala and P. Martínez. 2021. Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. *BMC Bioinformatics*, 22(Suppl 1):601.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, Seattle, WA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Manh Tran-Tien, Huu-Loi Le, Dang Nhat Minh, T. Tran Khang, Huy-The Vu, and Nguyen Minh-Tien. 2023. ViPubmedDeBERTa: A pre-trained model for Vietnamese biomedical text. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 831–840, Hong Kong, China. Association for Computational Linguistics.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 named entity recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2146–2153, Online. Association for Computational Linguistics.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xin-Yu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. 2026. TrustJudge: Inconsistencies of LLM-as-a-Judge and How to Alleviate Them.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general Chinese embedding. *Preprint*, arXiv:2309.07597.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

These appendix sections provide additional details that support the main paper.

A JSON data format for label mapping

Full-abstract prompting caused ambiguity when the same entity surface form appeared multiple times across sentences or when an entity was translated differently depending on local context. We therefore perform annotation mapping at the sentence level. For each English-Vietnamese sentence pair, the input JSON contains the English sentence, the post-edited Vietnamese translation, and the list of English entity spans with their UMLS semantic type labels. The Vietnamese annotation fields are left blank and filled by the LLM during projection.

This format allows the projection model to focus on local sentence-level alignment while preserving the original English entity labels. It also makes the projected spans easy to inspect in Label Studio, where annotators can verify the Vietnamese entity text, start and end offsets, and semantic type. Figure 3 shows an example of the JSON format used for label mapping.

B Human verification and adjudication details

B.1 Projection quality evaluation

Projection quality was evaluated using four criteria: text match, start index, end index, and exact match. Text match checks whether the projected Vietnamese mention matches the intended biomedical entity. Start and end indices assess whether the projected character-level span boundaries are correct. Exact match requires the entity text and both boundaries to be correct.

Inter-annotator agreement was computed at the sentence level. For each of the 450 mini-test sentences, annotators judged whether an error occurred under each projection criterion. The overall pooled κ was computed by aggregating all paired ratings across the four criteria.

Based on the consensus audit, entity-level projection accuracy over the 2,085 mentions reached:

- 98.32% for text match,
- 99.04% for start index,
- 98.27% for end index, and
- 98.32% for exact match before correction.

Metric	N	Agreement	P_o	P_e	κ
Text match accuracy	450	368/450	0.8178	0.1234	0.7921
Start index accuracy	450	393/450	0.8733	0.1225	0.8557
End index accuracy	450	406/450	0.9022	0.1227	0.8885
Exact match accuracy	450	367/450	0.8156	0.1236	0.7896
Overall pooled	**1800**	**1534/1800**	**0.8522**	**0.1230**	**0.8315**

Table 4: Inter-annotator agreement between two annotators measured by Cohen's kappa across four evaluation metrics. P_o denotes observed agreement and P_e denotes expected agreement by chance. The overall pooled result is computed by aggregating all paired ratings across the four metrics.

The 95% Wilson confidence interval for exact-match accuracy is 97.13%-99.51% (Wilson, 1927; Brown et al., 2001). If this estimate generalizes to the full test set, this corresponds to approximately 44-256 potentially erroneous entity pairs before correction.

B.2 Inter-annotator agreement details

Table 4 shows the agreement between two annotators during the quality check of projected Vietnamese annotations. The overall Cohen's κ score is 0.8315, which indicates strong agreement. Among the four metrics, end-index accuracy has the highest agreement, while exact-match accuracy is lower because it requires the entity text and both span boundaries to be correct. This suggests that most annotation disagreements come from difficult boundary or text-matching cases rather than random annotation noise.

B.3 Manual data checking with Label Studio example

Figure 4 shows a screenshot of the Label Studio interface used for manual verification. The interface displays each English sentence and its post-edited Vietnamese translation side by side, together with the projected biomedical entity labels. Annotators used this view to check whether the Vietnamese spans correctly matched the English source entities, whether span boundaries were accurate, and whether the UMLS semantic types were preserved. This step helped identify projection errors such as incorrect boundaries, untranslated abbreviations, and inconsistent translations of the same biomedical entity. The label panel also supports toggling the visibility of individual label pairs, helping annotators focus on specific entities during verification.

```
{
  "id": 1,
  "data": {
    "en_text": "DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection
    "vi_text": "Đột biến gen DCTN4: yếu tố ảnh hưởng đến nhiễm trùng phế khuẩn Pseudomonas aeruginosa mạn tính ở bệnh xơ nang phổi"
  },
  "annotations": [
    {
      "id": 1,
      "result": [
        {
          "id": "en_abs0001_sent01_ent01",
          "from_name": "en_label",
          "to_name": "en_text",
          "type": "labels",
          "value": {
            "start": 0,
            "end": 5,
            "text": "DCTN4",
            "labels": [
              "T103"
            ]
          }
        },
        {
          "id": "vi_abs0001_sent01_ent01",
          "from_name": "vi_label",
          "to_name": "vi_text",
          "type": "labels",
          "value": {
            "start": null,
            "end": null,
            "text": "",
            "labels": [
              "T103"
            ]
          }
        },
        {
          "id": "vi_abs0001_sent01_ent02",
```


Figure 3: Example of JSON data format for label mapping.

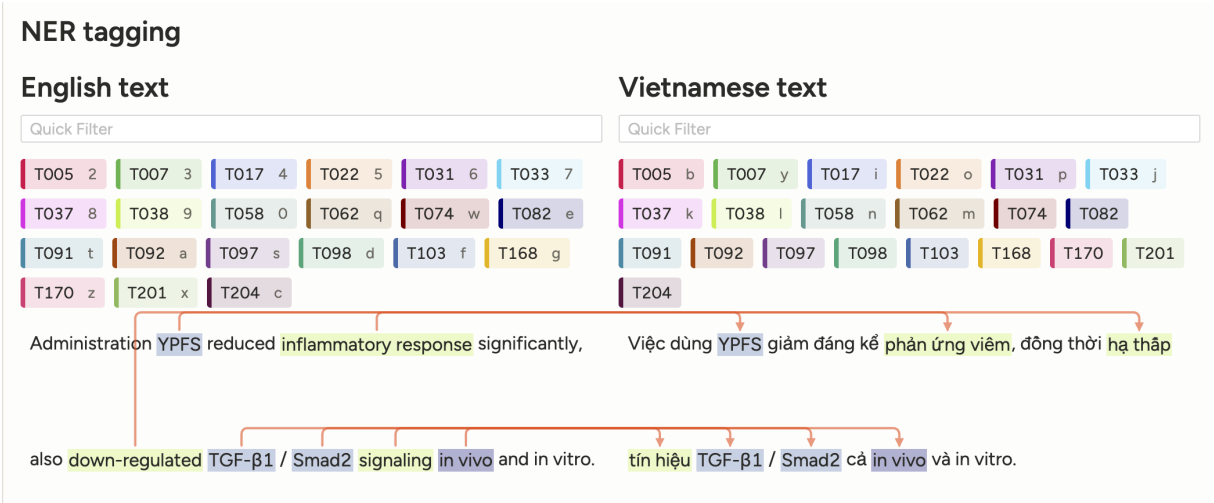

Figure 4: Screenshot of manual data checking in Label Studio.

C Details of Track 2.1 experimental settings

Figure 5 summarizes the Track 2.1 settings. We separate three component abilities: plain sentence translation, monolingual NER tagging, and tag-preserving translation.

Trans. only is a translation-only fine-tuned model trained to map a plain English sentence to a plain Vietnamese sentence, without generating NER tags.

NER only is a monolingual NER fine-tuned model trained to insert entity tags into a plain sentence in the same language.

Tag-aware Trans. is a tag-aware translation model trained to translate an already tagged English sentence into a tagged Vietnamese sentence while preserving entity labels.

M1 directly generates a tagged Vietnamese sentence from a plain English input, whereas **M2** and **M3** represent two cascaded alternatives: translate-then-tag and tag-then-translate.

This setup allows us to compare whether cross-lingual biomedical NER is better handled as a joint generation task or as a pipeline of separately optimized components.

D Implementation Details

Encoder fine-tuning. We fine-tune all parameters of PhoBERT, XLM-R, ViHealthBERT, and ViPubMedDeBERTa for token-level sequence labeling. We fine-tune the models for 3 and 5 epochs with learning rate selected from [2e-5, 3e-5, 5e-5], batch size 16, maximum sequence length 256, and warmup ratio 0.06, using FP16 mixed-precision

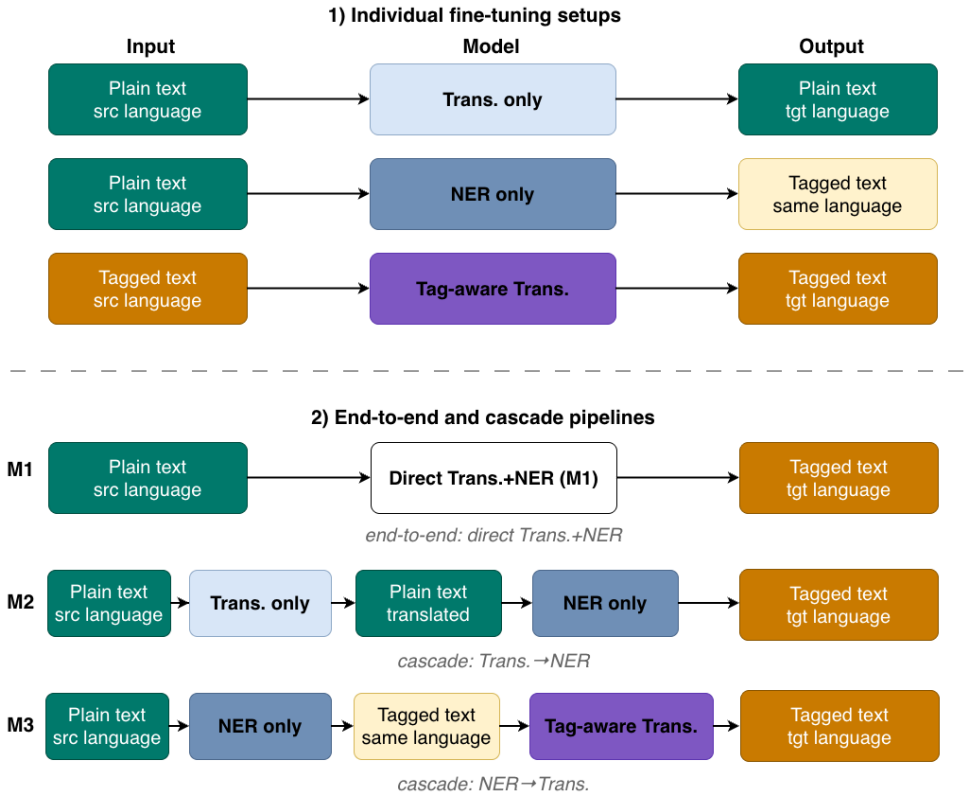

Figure 5: Overview of Track 2.1 settings. The figure compares single-task components with direct and cascaded cross-lingual NER pipelines.*Note*: Colors are used as secondary cues to distinguish plain text, tagged text, and model components; all boxes are directly labeled, so the figure remains readable in grayscale.

training. The best checkpoint is selected by validation F1 score. All experiments use random seed 42.

Encoder-decoder fine-tuning. We formulate the umT5 and NLLB experiments as text-to-text generation tasks with inline entity tags. We fine-tune the models for 5 epochs with learning rate 3e-5, batch size 16, gradient accumulation 2, maximum source/target length 256, warmup ratio 0.04, and greedy decoding. Models are evaluated and saved every 400 steps, with early stopping patience 3.

For Trans. only models, the best checkpoint is selected by validation BLEU.

For NER only models, the best checkpoint is selected by validation entity-level F1.

For M1, M2, and M3 models, we select checkpoints by validation BLEU and report NER F1 only in the final evaluation to avoid using judge-based evaluation during model selection.

Prompting setup. In prompt-based LLM experiments, we evaluate zero-shot, 1-shot, and 3-shot prompting using the prompt templates in Appendix H. For few-shot prompting, retrieved examples are selected from the training split only.

Vietnamese-input/Vietnamese-output NER uses Vietnamese demonstrations retrieved with Vietnamese_Embedding (Nguyen et al., 2025).

English-input/Vietnamese-output NER uses parallel English-Vietnamese demonstrations retrieved by English-side dense similarity with bge-large-en-v1.5 (Xiao et al., 2023).

The same number of demonstrations is used for all models under the same setting, and retrieved examples are inserted before the test input using the same JSON output format required from the model.

Open-source LLMs. For the zero-shot and few-shot experiments, we use Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen-SEA-LION-v4-8B-VL. We use temperature 0.1, maximum output length 2048.

Closed-source LLMs. For gpt-5.4 and gemini-3.1-flash-lite-preview, we use temperature 1.0, top-p 0.95, and a maximum output length of 2,048 tokens. We set the reasoning/thinking level to none for gpt-5.4 and minimal for gemini-3.1-flash-lite-preview. For both models, we allow up to two retries with a 3-second interval for transient

Tag	train	dev	test	Full Corpus
T005	702	223	172	1,097
T007	1,127	473	449	2,049
T017	12,588	4,107	3,772	20,467
T022	300	105	90	495
T031	772	269	212	1,253
T033	9,840	3,151	3,204	16,195
T037	1,058	435	357	1,850
T038	25,046	8,210	8,099	41,355
T058	14,808	4,680	4,778	24,266
T062	5,449	1,862	1,845	9,156
T074	1,169	493	355	2,017
T082	7,515	2,559	2,402	12,476
T091	521	198	196	915
T092	1,307	453	382	2,142
T097	1,051	365	360	1,776
T098	3,577	1,303	1,261	6,141
T103	22,436	7,499	7,400	37,335
T168	785	245	322	1,352
T170	5,999	1,948	2,359	10,306
T201	1,046	404	323	1,773
T204	4,921	1,862	1,750	8,533
Total	**122,017**	**40,844**	**40,088**	**202,949**

Table 5: Distribution of the 21 entity tags in the En-ViMedNER corpus across the train, development, test splits, and the full corpus.

API failures.

Hardware. All fine-tuning experiments were run on two NVIDIA L40 48GB GPUs.

E Other results

This section reports additional results that support the main findings in the paper.

Table 5 shows that the corpus is not balanced across labels. Some tags, such as T038, T103, T058, and T017, appear much more often than others, while T022 and T091 are relatively rare. This imbalance reflects the natural distribution of biomedical concepts in MedMentions, making the task more realistic. However, it also means that models may learn frequent entity types more easily than rare ones.

Table 6 provides detailed Vietnamese biomedical NER results on the full Vietnamese test set. The results show that model size alone does not determine performance. For example, ViPubMedDeBERTa-base achieves the best Micro-F1 despite being substantially smaller than XLM-R-large, while ViPubMedDeBERTa-xsmall remains competitive with much larger general-purpose encoders. This suggests that task- and domain-specific pretraining is more important than parameter count for Vietnamese biomedical NER.

The comparison between Vietnamese-supervised and English-supervised settings further highlights the value of En-ViMedNER. Although multilingual encoders can transfer across languages to some extent, their performance remains far below models fine-tuned on Vietnamese annotations. This gap indicates that English supervision and multilingual pretraining cannot fully capture Vietnamese biomedical terminology, word segmentation, and entity-boundary patterns.

Open-source LLM prompting performs substantially worse than fine-tuned encoder models, even though these LLMs are much larger. This finding reinforces that En-ViMedNER is useful not only as an evaluation benchmark but also as supervised training data for developing task-specific Vietnamese biomedical NER systems.

F Handling conflicts for BIO Tagging dataset

To convert the original entity annotations into token-level BIO labels, we first normalize span conflicts so that each token receives exactly one label. For nested entities, we keep only the outermost span, since it provides the most complete mention boundary for sequence labeling.

For non-nested overlapping entities with different semantic labels, we do not automatically resolve the conflict because choosing one label would introduce additional ambiguity into the gold annotations. Instead, we log the corresponding file path and exclude the sentence from the current experimental setting. In total, this filtering removes 128 cases across the train, development, and test splits, accounting for less than 0.3% of the full dataset.

G Postprocessing for LLMs experiments

For LLM-based experiments, we apply a lightweight postprocessing step before evaluation to improve the validity and consistency of predicted BIO sequences. This step is used only for LLM prompting outputs, since LLMs generate free-form responses and may produce malformed JSON, missing fields, or invalid BIO transitions. We do not apply the same postprocessing to fine-tuned encoder models, which are trained and decoded directly as token-level sequence taggers over a fixed BIO label space. Applying extra rule-based or LLM-based corrections to these models would introduce additional inference-time supervision, making the comparison less direct.

We keep "các" and "những" inside an entity span when adjacent words are labeled as part of the en-

Model	Params	Pretrain Data	Tokenization	Max Len	Precision	Recall	Micro-F1	Macro-F1
Vietnamese-supervised setting								
PhoBERT-base	135M	20GB	Word	256	52.65	49.98	51.28	47.22
PhoBERT-base-v2	135M	140GB	Word	256	52.60	50.42	51.49	46.48
PhoBERT-large	370M	20GB	Word	256	52.19	50.84	51.50	48.89
XLM-R-base	270M	2.5TB	Syllable	256	50.68	48.92	49.78	46.27
XLM-R-large	550M	2.5TB	Syllable	256	52.66	50.29	51.45	48.80
ViHealthBERT-word	135M	20GB+	Word	256	52.72	50.51	51.59	47.70
ViHealthBERT-syllable	135M	20GB+	Syllable	256	50.42	48.47	49.42	46.71
ViPubMedDeBERTa-xsmall	22M	158GB	Word	256	52.35	52.74	52.54	48.67
ViPubMedDeBERTa-base	86M	158GB	Word	256	52.86	52.54	**52.70**	**49.01**
English-supervised setting								
XLM-R-base	270M	2.5TB	Syllable	256	27.80	35.48	31.12	27.10
XLM-R-large	550M	2.5TB	Syllable	256	33.54	40.13	36.54	34.21
Open-source LLMs								
Qwen2.5-3B-Instruct-0shot	3B	18T	BPE	2048	3.44	0.85	1.37	0.95
Qwen2.5-3B-Instruct-1shot	3B	18T	BPE	2048	8.87	2.48	3.87	3.09
Qwen2.5-3B-Instruct-3shot	3B	18T	BPE	2048	11.04	2.66	4.29	3.29
Qwen2.5-7B-Instruct-0shot	7B	18T	BPE	2048	2.66	0.33	0.59	0.67
Qwen2.5-7B-Instruct-1shot	7B	18T	BPE	2048	10.26	3.04	4.69	3.66
Qwen2.5-7B-Instruct-3shot	7B	18T	BPE	2048	13.30	3.16	5.10	4.22
Qwen-SEA-LION-v4-8B-VL-0shot	8B	9M SFT pairs	BPE	2048	10.57	1.57	2.73	2.46
Qwen-SEA-LION-v4-8B-VL-1shot	8B	9M SFT pairs	BPE	2048	14.88	6.13	8.68	7.17
Qwen-SEA-LION-v4-8B-VL-3shot	8B	9M SFT pairs	BPE	2048	19.84	9.50	12.85	10.08

Table 6: Detailed Vietnamese biomedical NER results on the full test set with Precision, Recall, Micro-F1, and Macro-F1. Encoder-based models are evaluated under Vietnamese-supervised and English-supervised settings, while open-source LLM prompting results are additionally reported on the full test set. Closed-source LLMs are omitted from this table because they are evaluated only on the mini-test set (due to API cost). *Note:* For Qwen2.5 models, "Pretrain Data" denotes the reported 18T-token pretraining corpus of the Qwen2.5 series. For Qwen-SEA-LION-v4-8B-VL, "Pretrain Data" denotes the approximately 9M supervised fine-tuning instruction-text pairs used for SEA-LION regional adaptation.

tity, but remove "khác" and "này" because they usually act as modifiers rather than core entity tokens. We also remove "mỗi" unless the corresponding English entity begins with per, to avoid incorrectly including mistranslated rate or frequency expressions.

Finally, we repair invalid BIO transitions such as O-I and B-X I-Y. For simple O-I cases, the invalid I tag is converted into a valid beginning tag when appropriate. For inconsistent adjacent entity types, we use the prompt template in Figure 6 to select the most plausible entity type from the sentence context and tag definitions.

Invalid outputs. For LLM prompting experiments on Task 2: cross-lingual NER, an output is marked as invalid if it cannot be parsed into the required JSON/tagged format or if it contains no valid entity tags (all tokens are tagged as "O"). Invalid outputs are not discarded. When no valid predicted entity can be extracted, the prediction set is treated as empty: all gold entities are counted as FN, and no predicted entities are counted as TP. The number of invalid outputs is reported separately as **Inv.**.

H Prompt templates

This appendix presents the prompt templates used in this study.

Prompt template for BIO conflict resolution

```
You are a biomedical named entity expert.

# TASK
You are given a biomedical BIO tagging conflict. Resolve the conflict.

# INPUT

Sentence:
{sentence}

Tokens:
{json.dumps(tokens, ensure_ascii=False)}

Current labels:
{json.dumps(labels, ensure_ascii=False)}

Conflict:
- token[{i}] = "{tokens[i]}" has label "{labels[i]}"
- token[{j}] = "{tokens[j]}" has label "{labels[j]}"
This is invalid because B-X cannot be followed immediately by I-Y when X != Y.

Choose the single best entity type for this local span based on the sentence context.

# TAG DEFINITIONS
{tag_legend}

Return JSON only:
{{
  "resolved_type": "<one tag type such as T033>",
  "tokens": ["<token1>", "<token2>", "<token3>"],
  "labels": ["<label1>", "<label2>", "<label3>"]
}}
```

Figure 6: Prompt template used for resolving invalid BIO transitions in biomedical named entity recognition outputs.

Prompt template for few-shot Vietnamese NER experiments

You are a biomedical named entity recognition expert.

TASK
You are given one Vietnamese biomedical sentence. Perform token-level biomedical named entity recognition using the BIO tagging scheme.

Your job is to:
1. split the sentence into tokens,
2. assign exactly one label to each token,
3. use BIO tagging:
- B-<TAG> for the beginning of an entity,
- I-<TAG> for the continuation of the same entity,
- O for a token outside any entity.

IMPORTANT RULES
- Every token must receive exactly one label.
- Preserve the original sentence content.
- Do not add, remove, paraphrase, or translate words.
- Use only the semantic tags defined below.
- If a token is not part of a biomedical named entity, label it O.
- If an entity has multiple tokens, the first token must be B-<TAG> and the remaining tokens must be I-<TAG>.
- Do not create overlapping entities.
- Learn the JSON output style from the examples below.

TAG DEFINITIONS
{tag_legend}

ANSWER FORMAT
Return JSON only:
{{
  "vi_text": "<original Vietnamese sentence>",
  "tokens": ["<token1>", "<token2>", "<token3>"],
  "labels": ["<label1>", "<label2>", "<label3>"]
}}

Constraints for output JSON:
- "tokens" and "labels" must have the same number of items.
- Each label in "labels" must be exactly one of: O, B-<TAG>, or I-<TAG> where <TAG> is from TAG DEFINITIONS.
- Keep "vi_text" unchanged from the input sentence.

FEW-SHOT EXAMPLES
{example_text}

INPUT SENTENCE
{input_text}

Figure 7: Prompt template used for few-shot Vietnamese NER experiments. For zero-shot experiments, the section **# FEW-SHOT EXAMPLES** is excluded.

```
## EXAMPLE 1
Vietnamese sentence:
Đột biến gen DCTN4: yếu tố ảnh hưởng đến nhiễm trùng phế khuẩn Pseudomonas aeruginosa mạn tính ở bệnh xơ
 nang phổi

JSON answer:
{
  "vi_text": "Đột biến gen DCTN4: yếu tố ảnh hưởng đến nhiễm trùng phế khuẩn Pseudomonas aeruginosa mạn
tính ở bệnh xơ nang phổi",
  "tokens": ["Đột", "biến", "gen", "DCTN4", "yếu", "tố", "ảnh", "hưởng", "đến", "nhiễm", "trùng", "phế",
 "khuẩn", "Pseudomonas", "aeruginosa", "mạn", "tính", "ở", "bệnh", "xơ", "nang", "phổi"],
  "labels": ["O", "O", "O", "B-T103", "O", "O", "O", "O", "O", "B-T038", "I-T038", "I-T038", "I-T038", "
I-T038", "I-T038", "I-T038", "I-T038", "O", "B-T038", "I-T038", "I-T038", "I-T038"]
}
```

Figure 8: Example of **example_text** in one-shot Vietnamese NER prompt.

Prompt template for few-shot cross-lingual NER experiments

You are a biomedical named entity recognition expert for English-to-Vietnamese transfer.

TASK
Your job is to:
1. translate the input sentence into Vietnamese in medical domain with medical terms,
2. split the Vietnamese sentence into tokens,
3. assign exactly one BIO label to each Vietnamese token.
4. Use BIO tagging:
- B-<TAG> for the beginning of an entity,
- I-<TAG> for the continuation of the same entity,
- O for a token outside any entity.

IMPORTANT RULES
- The output sentence must be in Vietnamese.
- The meaning must stay faithful to the English input.
- The tokens and labels must correspond to the Vietnamese sentence only.
- Every token must receive exactly one label.
- Use only the semantic tags defined below.
- If a token is not part of a biomedical named entity, label it O.
- If an entity has multiple tokens, the first token must be B-<TAG> and the remaining tokens must be I-<TAG>.
- Do not create overlapping entities.

TAG DEFINITIONS
{tag_legend}

ANSWER FORMAT
Return JSON only:
{{
  "vi_text": "<Vietnamese translation>",
  "tokens": ["<token1>", "<token2>", "<token3>"],
  "labels": ["<label1>", "<label2>", "<label3>"]
}}

FEW-SHOT EXAMPLES
{example_text}

INPUT ENGLISH SENTENCE
{input_text}

Figure 9: Prompt template used for few-shot cross-lingual NER experiments. For zero-shot experiments, the section **# FEW-SHOT EXAMPLES** is excluded.

```
## EXAMPLE 1
English sentence:
DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis

JSON answer:
{
  "vi_text": "Đột biến gen DCTN4: yếu tố ảnh hưởng đến nhiễm trùng phế khuẩn Pseudomonas aeruginosa mạn
tính ở bệnh xơ nang phổi",
  "tokens": ["Đột", "biến", "gen", "DCTN4", "yếu", "tố", "ảnh", "hưởng", "đến", "nhiễm", "trùng", "phế",
 "khuẩn", "Pseudomonas", "aeruginosa", "mạn", "tính", "ở", "bệnh", "xơ", "nang", "phổi"],
  "labels": ["O", "O", "O", "B-T103", "O", "O", "O", "O", "O", "B-T038", "I-T038", "I-T038", "I-T038", "
I-T038", "I-T038", "I-T038", "I-T038", "O", "B-T038", "I-T038", "I-T038", "I-T038"]
}
```

Figure 10: Example of **example_text** in one-shot cross-lingual NER prompt.

Prompt template for LLM-as-a-judge evaluation of cross-lingual NER outputs

```
Given the prediction tagged:
"{pred_sent}"

Prediction tagged entities list:
"{pred_entities_list}"

Given ground truth:
"{gt_sent}"

Ground truth entities list:
"{gt_entities_list}"

For each ground truth entity, find the best matching predicted entity, allowing paraphrases. Each predicted
    entity can be used at most once.

Rules:
- "pred text label" = "correct" if pred text matches gt meaning; otherwise "incorrect"; use "miss" if no
    match.
- "pred tag label" = "correct" if pred tag = gt tag; otherwise "incorrect"; use "miss" if no match.
- If no match, set "pred text" = "" and "pred tag" = "".
- Return unmatched predicted entities separately.

Return ONLY valid JSON parseable by json.loads() as:
{
  "matches": [
   {
    "gt text": "...",
    "gt tag": "...",
    "pred text": "...",
    "pred text label": "correct|incorrect|miss",
    "pred tag": "...",
    "pred tag label": "correct|incorrect|miss"
   }
  ],
  "unmatched": [
   {
    "pred text": "...",
    "pred tag": "..."
   }
  ]
}

Do not output markdown, code fences, explanations, or any text outside the JSON object.
```

Figure 11: Prompt template used for LLM-as-a-judge evaluation of cross-lingual NER outputs.