Assessing Prompt Design Choices in Biomedical Information Extraction for Low-Resource Languages

Anonymous ACL submission

Abstract

Despite the strong multilingual abilities of modern LLMs, biomedical information extraction remains inconsistent for low-resource, morphologically rich languages such as Bangla/Bengali and Basque. Prior investigations of prompt design and output schemas have been focused on high-resource settings. To bridge this gap, we systematically evaluate biomedical named entity recognition (NER) with open LLMs under multiple prompting settings. We find that span-based extraction is much more effective than BIO tagging for LLM prompting across all languages, while moving from statement-based prompting to question-based prompting has a stronger effect on low-resource languages than on high-resource languages (e.g., +57% for Bangla and +109% for Basque, but only +28% for English and +22% for Spanish). Our breakdowns by error type show that translation-based prompting cuts Bangla hallucinations by 64% and QA-style prompting lowers Basque empty prediction errors by 61%. Our results offer practical guidance for building reliable multilingual biomedical NER systems in low-resource languages. We release our code at https://github.com/<anonymized>.

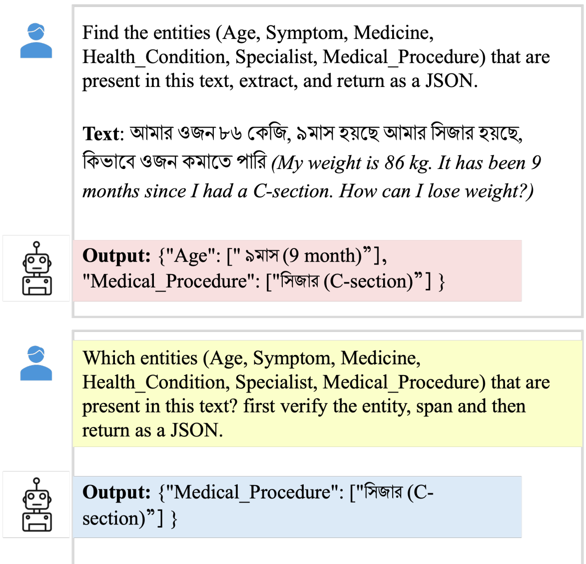

Figure 1: A statement-based prompt for biomedical NER in Bangla results in an incorrect extraction (red), while a question-based prompt results in a correct extraction (blue).

1 Introduction

Research on health and medical texts is important because it affects people's lives, can reduce cost and workload, and supports better clinical decisions (Agrawal et al., 2022; Averly and Ning, 2025). However, most medical information appears in unstructured text, so medical named entity recognition (NER) is needed to identify and label key medical terms (Averly and Ning, 2025). Creating these labels requires medical expertise, which makes annotation slow, expensive, and hard to scale. Traditional fine-tuning for NER also needs many labeled examples. To reduce this dependence, prior work has explored prompting and in-context learning with large language models (LLM) achieving great performance in high-resource languages like English (Ashok and Lipton, 2023; Li et al., 2023a; Li and Zhang, 2024; Wang et al., 2025; Averly and Ning, 2025). In contrast, for low-resource languages, we still lack a systematic understanding of how different prompting strategies affect multilingual LLM performance in biomedical NER (Azime et al., 2025; Kumar et al., 2025).

We present a study of how different prompting strategies behave in Bangla/Bengali and Basque, two morphologically rich languages (Bhattacharyya and Bhattacharya, 2025; López de Lacalle et al., 2020) with complex grammar that pose challenging low-resource settings. To test whether prompting effects generalize across languages, we also evaluate Spanish and English as high-resource comparisons, which lets us separate shared prompting behavior from language-specific effects under the same multilingual models. We compare two output representations, Beginning–Inside–Outside (BIO) tagging (Chen et al., 2022) and span-based extraction, and provide practical guidance for converting BIO outputs into span format. Finally, we analyze which prompting strategies help reduce common extraction errors, including entity hallucination, over-generation, and boundary mistakes. We summarize these goals in three research questions:

- Which prompting strategy most reliably improves biomedical NER for low-resource languages?
- Does the span-based format beat BIO tagging format, and why?
- Which errors occur most often across prompts, and which prompting choices reduce hallucination, over-generation, and boundary errors?

We show that LLMs struggle generating outputs in BIO format and that span-based extraction is a much better option. We also find that formulating NER as a question-style task (see Figure 1) obtains the best results, providing major improvements for low-resource languages like Bangla and Basque. Our outcomes can benefit biomedical NER not only in other low-resource languages but also in other multilingual settings. Since multilingual models share parameters across languages, methods designed for low-resource settings often strengthen the shared multilingual backbone (Pfeiffer et al., 2020; Liu et al., 2021; Choenni et al., 2023; Pham et al., 2024).

2 Related Work

Prior work has explored zero-shot and few-shot methods for Bangla across several NLP tasks (e.g., Dementieva et al. 2025; Adak et al. 2025; Li et al. 2023b; Shafayat et al. 2024); for example, Hasan et al. (2024) use 3-shot and 5-shot prompting for sentiment analysis. Prompting has also been applied to general Bangla NER: Mahtab et al. (2025) provide an English instruction prompt describing the BIO tags and rules, include 10 in-context input–output examples, and then prompt the model to label a new Bangla sentence in BIO format. Early work on Bangla biomedical/telemedicine NER (e.g., Islam et al. 2022; Sazzed 2022) primarily focused on dataset construction. More recently, Khan et al. (2023) introduce Bangla-HealthNER and evaluate fine-tuned BanglaBERT, BanglishBERT, and mBERT models, and also report substantially lower performance for zero-shot ChatGPT than for supervised fine-tuning.

For Basque, there is prior work on biomedical NER, and most of it relies on supervised training or fine-tuning with BIO-style output formats. Urbizu et al. (2022) introduce the BasqueGLUE NLU benchmark; they use standard fine-tuning and report baseline results, and for the NER task they follow the BIO annotation scheme. Zanoli et al. (2024b) examine whether a multilingual clinical corpus is effective for disorder NER; they train and fine-tune supervised NER models and use IBO/BIO-style tagging.

However, existing work for both Bangla and Basque largely emphasizes supervised BIO tagging, and we still lack a clear picture of how prompt design and output format affect LLM-based biomedical NER in these languages. To fill this gap, we systematically evaluate multiple prompting strategies and compare BIO tagging with span-based JSON extraction.

3 Methods

Figure 2 details our NER evaluation pipeline and the different configurations we have explored in low-resource languages. We start from a base prompt that includes instructions for the task and some rules and hints to guide the LLM when annotating the entities. In all cases, the instructions are in English but indicate the target language. For example, "You are a biomedical NER assistant that performs Named Entity Recognition in Bangla." Building on this base, we explore different configurations involving the output format of the annotations, whether or not to include demonstration examples, the format of these examples, or how the model should address the task.

3.1 Datasets

For our Bangla experiments, we use Bangla-HealthNER (Khan et al., 2023), a large Ben-

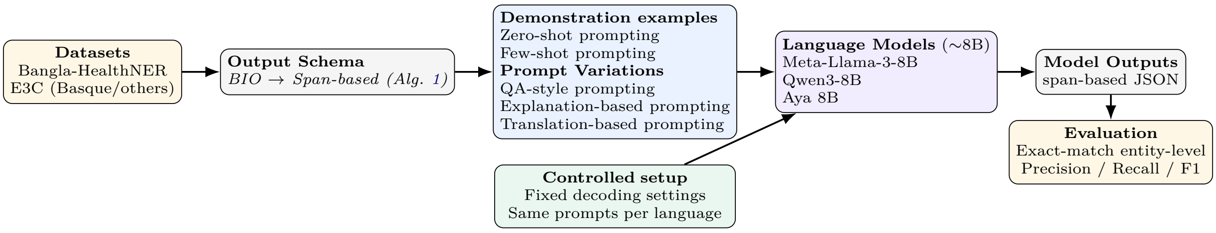

Figure 2: Overview of our biomedical NER evaluation pipeline. We vary (i) output schema (BIO vs span-level JSON) and (ii) prompting strategy, keep inference settings fixed, and evaluate with exact-match entity-level F1.

gali biomedical NER dataset built from consumer health Q&A collected from a public online health platform in Bangladesh. It contains 31,783 samples and 144,136 sentences, annotated in token-level (BIO-style) format. The dataset includes seven entity types: Symptom, Health_Condition, Medicine, Specialist, Age, Dosage, and Medical_Procedure. We include Bangla-HealthNER because it is one of the largest publicly available human-annotated Bengali medical NER resources.

For Basque, Spanish, and English, we use the European Clinical Case Corpus (E3C; Magnini et al., 2020), a multilingual collection of clinical case narratives in five languages. The corpus reports clinical-case counts/tokens of English 9,533 (928K), Spanish 1,400 (531K), and Basque 122 (26K). Annotations are provided in a two-level scheme: a THYME-style layer (Styler et al., 2014) that covers time expressions events and actors, and a clinical entity layer that provides UMLS[1] Concept Unique Identifiers of different clinical disorders. From the THYME layer, we keep entities corresponding to the Patient and H-Professional (health professionals) categories. From the UMLS layer, we following previous work (Zanoli et al., 2024a) and label all disorders with the same Disorder category.

3.2 Output Schema

For all datasets, we evaluate two representations: (1) the original BIO token-level labels and (2) a span-based format (entity mentions as text spans), allowing direct comparison between tag-based and span-based extraction.

BIO format (bio) is a token-level tagging scheme for NER (Ramshaw and Marcus, 1995) that has been previously used with LLMs for Bangla NER (Mahtab et al., 2025). Given a

```
SYSTEM_PROMPT = (
You are a biomedical NER assistant. Identify
ONLY entities explicitly present in the sentence
and output STRICTLY in BIO format.
Allowed output labels (BIO):
B-Age I-Age
B-Dosage I-Dosage
B-Symptom I-Symptom
B-Medicine I-Medicine
B-Health_Condition I-Health_Condition
B-Specialist I-Specialist
B-Medical_Procedure I-Medical_Procedure
O
For the text Output exactly one label per token,
space-separated.
BIO_PROMPT = (
"Text: \"আমার পিত্ত থলিতে পাথর আছে। গত ২.৫ বছর
যাবৎ এটা হয়েছে। যার আকার ৪ সেমি। এতদিন তীব্র
কোন ব্যাথা ছিল না কিন্তু গত ৮ দিনের মধ্যে ৩ দিন
পেটে তীব্র ব্যাথা হয়েছিল এবং পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়। পিত্ত থলিতে পাথর
হলে কী পেট ব্যাথার সাথে সাথে পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়? এবং কেন এই চাপ
দিলে ব্যাথা অনুভূত হয়?? অভিজ্ঞ ডাক্তারের পরামর্শ
চাচ্ছি।\"\n”

Answer: O O B-Health_Condition I-
Health_Condition O O
O O O O O O O
O O O O O
O O O B-Symptom O O O O O O O O O B-
Symptom I-Symptom O O O O O O O O B-
Symptom I-Symptom O O
O B-Specialist O O O
```

Figure 3: Illustration of a BIO prompt for Bangla biomedical NER, where the system instruction enforces strict BIO tagging and the model outputs one label per token.

tokenized sentence $x = (x_1, \ldots, x_n)$, dataset assigns a tag sequence $y = (y_1, \ldots, y_n)$ where each tag is one of:

$$y_i \in \{O\} \cup \{B\text{-}t,\ I\text{-}t \mid t \in \mathcal{T}\},$$

$\mathcal{T}$ is the set of entity types (e.g., AGE, SYMPTOM, etc.). Here, B marks the beginning of an entity span, I marks inside the same entity span, and O means not an entity.” Figure 3 shows the prompt for the BIO tagging format.

[1] https://www.nlm.nih.gov/research/umls/

Algorithm 1 BIO $\rightarrow$ span-level JSON

Require: Tokens $x_{1:n}$, BIO tags $y_{1:n}$, entity types $\mathcal{T}$
Ensure: $\mathcal{J}$: map $t \in \mathcal{T} \mapsto$ list of extracted span strings

$\mathcal{J}(t) \leftarrow [\,]\ \forall t \in \mathcal{T}$
$i \leftarrow 1$
while $i \leq n$ **do**
  if $y_i = \texttt{B-}t$ for some $t \in \mathcal{T}$ **then**
    $s \leftarrow i$
    $i \leftarrow i + 1$
    while $i \leq n$ **and** $y_i = \texttt{I-}t$ **do**
      $i \leftarrow i + 1$
    end while
    $e \leftarrow i - 1$
    $span \leftarrow \textsc{Detok}(x_{s:e})$
    $span \leftarrow \textsc{Clean}(span)$
    if $span \neq \emptyset$ **then**
      $\mathcal{J}(t) \leftarrow \mathcal{J}(t) \,\|\, [span]$
    end if
  else
    $i \leftarrow i + 1$
  end if
end while
return $\mathcal{J}$

```
SYSTEM_PROMPT = (
You are a question-answering assistant that performs
medical Named Entity Recognition ..... )
QA-Style_PROMPT = (
“Question: Which entities (Age, Symptom, Medicine,
Health_Condition, Specialist, Medical_Procedure)
are present in this text?\n”
"Text: \"আমার বয়স ২৪ ওজন ৫৮ কেজি আমার কিছুখন
পর পর প্রসাব হয় খুবকম এখন আমি কি করতে
পারি। \"\n"
"Answer: {\"Age\":[], "
"\"Symptom\":[\"প্রসাবের রাস্তায়
জ্বালাপোড়া\",\"ব্যথা\",\"প্রসাবে মাঝে মাঝে গন্ধ\","
"\"Medicine\":[], "
"\"Health_Condition\":[], "
"\"Specialist\":[], "
"\"Medical_Procedure\":[]}\n\n” )
```

Figure 4: Illustration of a question-style prompt for Bangla biomedical NER, where the yellow-highlighted text explicitly asks the model to identify predefined medical entity types in the input text.

Span-based format (json) means the model does not output a tag for every token. Instead, it directly outputs the entity mentions as text spans from the sentence, usually grouped by entity type in a JSON-like structure. Many prior works adopt span-based NER models (Sohrab and Miwa, 2018; Rojas et al., 2022). In contrast, our approach predicts BIO tags and then performs BIO decoding (chunk extraction) to convert token-level predictions into span strings. Formally, BIO decoding yields a set of extracted spans:

$$\hat{\mathcal{S}} = \left\{ (t, x_{s:e}) \;\middle|\; \begin{array}{l} y_s = \texttt{B-}t,\ \forall i \in (s{+}1, \ldots, e) : y_i = \texttt{I-}t, \\ y_{e+1} \neq \texttt{I-}t \end{array} \right\}$$

where $\mathcal{T}$ is the set of entity types and $x_{s:e}$ is the token span reconstructed into text. We used Algorithm 1 to detokenize each extracted span and append it to $\mathcal{J}(t)$. Code is also available at: `https://github.com/<anonymized>`.

3.3 Demonstration Examples

Zero-shot prompting (zero) is a strategy where we provide only the task instruction and the list of entity types, without any labeled examples (Liu et al., 2023). We provided detailed guidelines to the model, e.g., instructing it to skip negated symptoms, not confuse duration with age, not treat lab/test names as symptoms, and to select short, clean spans with no duplicates. This setting tests whether the model can perform biomedical extraction with only instructions and no demonstrations.

Few-shot prompting (few) is a way to use an LLM without training it, where we show the model a small number of labeled examples (Pan et al., 2023; Cheng et al., 2025) inside the prompt and then ask it to do the same task for a new input. Such labeled examples are known as the *support set*, $\mathcal{S}_K$, where K labeled examples are included in the prompt. Formally, we combine the base prompt P_{base} with this support set, and find the LLM's most probable output, $\hat{o}_{\text{FS}}$, given this combined prompt:

$$P_{\text{FS}} = P_{\text{base}} \oplus \mathcal{S}_K$$
$$\hat{o}_{\text{FS}} = \arg\max_{o} p(o \mid x, P_{\text{FS}})$$

The few-shot prompt contains $K = 9$ in-context examples in the target language. We designed the examples in a way that each entity type appears at least once, and the set includes both short and long examples.

3.4 Prompt Variations

Question-style prompting (qa) appends an explicit question q to the prompt while using the same support set $\mathcal{S}_K$:

$$P_{\text{QA}} = P_{\text{base}} \oplus \mathcal{S}_K \oplus q$$
$$\hat{o}_{\text{QA}} = \arg\max_{o} p(o \mid x, P_{\text{QA}})$$

This approach presents the biomedical entity extraction task as a direct question, such as “Which entities of types (Age, Symptom, …) appear in this text?””, as shown in Figure 4.

```
SYSTEM_PROMPT = (
You are a biomedical NER assistant that performs
medical Named Entity Recognition ..... )

Explanation-based_PROMPT = (

"Text: \"আমার পিত্ত থলিতে পাথর আছে। গত ২.৫ বছর
যাবৎ এটা হয়েছে। যার আকার ৪ সেমি। এতদিন তীব্র
কোন ব্যাথা ছিল না কিন্তু গত ৮ দিনের মধ্যে ৩ দিন
পেটে তীব্র ব্যাথা হয়েছিল এবং পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়। পিত্ত থলিতে পাথর
হলে কী পেট ব্যাথার সাথে সাথে পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়? এবং কেন এই চাপ
দিলে ব্যাথা অনুভূত হয়?? অভিজ্ঞ ডাক্তারের পরামর্শ
চাচ্ছি। \"\n”

"Description of text: The Bangla text says the
patient has gallbladder stones for about 2.5 years
and the stone size is 4 cm. The phrase “পিত্ত থলিতে
পাথর” is a Health_Condition entity (a
diagnosis/condition). Because of this condition,
the patient reports pain-related symptoms, such
as severe abdominal pain and pain when pressing
under the right ribs (these are Symptom
entities).\n"

"Answer: {\"Age\":[], "
"\"Symptom\":[\"ব্যাথা\",\"পেটে তীব্র ব্যাথা\",\"চাপ
দিলে ব্যাথা\",\"পেট ব্যাথার\"], "
"\"Medicine\":[], "
"\"Health_Condition\":[\"পিত্ত থলিতে পাথর\"], "
"\"Specialist\":[], "
"\"Medical_Procedure\":[]}\n\n"
```

Figure 5: Illustration of an explanation-based prompt for Bangla biomedical NER, where the yellow-highlighted text provides a natural-language explanation of the clinical context and entity semantics.

```
SYSTEM_PROMPT = (
You are a biomedical NER assistant that performs
medical Named Entity Recognition ..... )

Translation-based_PROMPT = (

"Text: \"আমার পিত্ত থলিতে পাথর আছে। গত ২.৫ বছর
যাবৎ এটা হয়েছে। যার আকার ৪ সেমি। এতদিন তীব্র
কোন ব্যাথা ছিল না কিন্তু গত ৮ দিনের মধ্যে ৩ দিন
পেটে তীব্র ব্যাথা হয়েছিল এবং পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়। পিত্ত থলিতে পাথর
হলে কী পেট ব্যাথার সাথে সাথে পেটের ডানদিকে পাজরের
নিচে চাপ দিলে ব্যাথা অনুভূত হয়? এবং কেন এই চাপ
দিলে ব্যাথা অনুভূত হয়?? অভিজ্ঞ ডাক্তারের পরামর্শ
চাচ্ছি। \"\n”

”Translation of the text: I have gallstones in my
gallbladder. This has been happening for the last
2.5 years. The stone size is 4 cm. I did not have any
severe pain for a long time, but within the last 8
days, I had severe stomach pain on 3 days, and I
feel pain when I press under the right rib on the
right side of my abdomen. If there are gallstones,
does abdominal pain occur along with pain when
pressing under the right rib on the right side? And
why do I feel pain when I press there?? I want
advice from an experienced doctor.\n"

"Answer: {\"Age\":[], "
"\"Symptom\":[\"ব্যাথা\",\"পেটে তীব্র ব্যাথা\",\"চাপ
দিলে ব্যাথা\",\"পেট ব্যাথার\"], "
"\"Medicine\":[], "
"\"Health_Condition\":[\"পিত্ত থলিতে পাথর\"], "
"\"Specialist\":[], "
"\"Medical_Procedure\":[]}\n\n"
```

Figure 6: Illustration of a translation-based prompt for Bangla biomedical NER, where the yellow-highlighted text shows the English translation of the original Bangla clinical text used for entity extraction.

Explanation-based prompting (expl) is a prompting strategy (Figure 5) that provides a brief description of what entities are likely present in the current text. First, the prompt explains the label boundaries (e.g., HEALTH_CONDITION is a diagnosis or stated condition, while SYMPTOM is a complaint). Then, for each example, we add a short natural-language description in English that highlights the relevant cues in the input. We have used explanation only for examples, not the entire prompt. This prompt uses the same support set $\mathcal{S}_K$ and adds label definitions $D = \{d_\ell\}_{\ell \in \mathcal{L}}$:

$$P_{\text{Desc}} = P_{\text{base}} \oplus \mathcal{S}_K \oplus D$$

$$\hat{o}_{\text{Desc}} = \arg\max_{o} \, p(o \mid x, P_{\text{Desc}})$$

Translation-based prompting (trans) is a process of prepending an English rendering (Figure 6) of the input while also keeping the original Bangla/Basque text in the prompt. This encourages the model to reason in English, which can help when the model is stronger in English than in the target language. However, this approach can fail when translation paraphrases the meaning, drops details, mistranslates medical terms, or changes span boundaries. Therefore, we treat translation-based prompting as a practical baseline rather than a guaranteed improvement. Let $T(\cdot)$ be a translation function (e.g., Bangla/Basque → English). We use the same support set $\mathcal{S}_K$, but the input includes the translated text:

$$x' = T(x)$$

$$P_{\text{Trans}} = P_{\text{base}} \oplus \mathcal{S}_K$$

$$\hat{o}_{\text{Trans}} = \arg\max_{o} \, p(o \mid x', P_{\text{Trans}})$$

3.5 Language Models

We primarily conduct our experiments using **Meta-Llama-3-8B** (Grattafiori et al., 2024), a widely adopted open-weights model with strong general performance. However, to test the effect of different model families, we also evaluate the best-performing prompt strategy under LLama-3 on two additional models: **Qwen3-8B** (Team, 2025) and **Aya** (Aryabumi et al., 2024). We include Qwen3 because it is a competitive recent model that often performs particularly well on high-resource languages such as English, and Aya because

Prompt	Bangla	Basque	Spanish	English
bio_few	0.054	0.104	0.019	0.120
json_few	**0.345**	**0.312**	**0.445**	**0.527**

Table 1: Biomedical NER Performance across output schemas with Llama3-8B. In both cases, we apply few-shot prompting.

Prompt	Bangla	Basque	Spanish	English
json_zero	0.291	0.241	0.403	0.422
json_few	0.345	0.312	0.496	0.527
json_few_qa	**0.458**	**0.503**	**0.528**	**0.541**
json_few_trans	0.347	0.441	0.494	0.313
json_few_expl	0.417	0.386	0.477	0.304

Table 2: Prompting Strategy Comparison for Multilingual Biomedical NER Using Meta-Llama-3-8B

it is designed for multilingual use and thus provides a useful contrast for low-resource settings. To ensure a fair comparison, we select similar size models (all in the ∼8B parameter range) and and keep inference settings fixed.

4 Experimental Results

4.1 Output Schema Comparison: BIO Tagging vs Span-Based

Table 1 shows a large gap between BIO tagging and span-based extraction, indicating that the output schema strongly affects LLM-based NER. Across languages, switching from *bio_few* to *json_few* yields consistent absolute F1 gains: +0.291 (Bangla; 0.054 → 0.345), +0.208 (Basque; 0.104 → 0.312), +0.426 (Spanish; 0.019 → 0.445), and +0.407 (English; 0.120 → 0.527). BIO tagging is likely harder for LLMs because the model must label every token in the exact order, with even one missing/extra token (or a tokenization mismatch in Bangla/Basque) breaking the whole alignment. Span-based extraction is easier because the model can simply copy the entity text spans and return them as clean JSON, which matches how LLMs naturally answer.

4.2 Prompting Strategy Comparison Across Languages

Table 2 shows that QA-style prompting performs best across all languages. It even outperforms detailed label explanations (*expl*) and translation of the input to English (*trans*), despite using only simple WH-questions. The strong performance of QA-style prompting is likely because modern LLMs are instruction tuned on many question-answer pairs. Prior work reformulating NER as machine reading comprehension similarly finds that QA-style formulations improve NER, attributing gains to query conditioning and (when available) semantically informative queries that encode entity-type knowledge (Li et al., 2020).

Prompting strategy has a stronger effect on low-resource languages than on high-resource languages. For example, moving from zero-shot to question-style prompting improves Bangla by 57.4% and Basque by 108.7%, while Spanish improves by 21.8% and English by only 28.2%.

4.3 Biomedical Entity-type-wise Comparison

Table 3 and Table 4 break down performance of prompting strategies by named entity types for Bangla and Basque, respectively. In both languages, question-style prompting is best for most types, with the largest absolute gains for MEDICAL_PROCEDURE in Bangla (37.8%) and H-PROFESSIONAL in Basque (37.4%), indicating that such prompting is especially helpful for rarer or harder-to-extract categories.

Occasionally, explanation or translation-based prompting outperforms question-based prompting. Explanation-based prompting helps Bangla SYMPTOMs and translation-based prompting helps Bangla SPECIALISTs. But the gains are modest over question-based prompting, and translation-based prompting fails badly for the AGE category.

4.4 Model Comparison Under the Best Prompt

Table 5 compares different LLMs under the best-performing (question-style) prompt. Overall, Llama-3-8B performs best on Bangla (F1 = 0.494), while Qwen3-8B achieves the highest score on English (F1 = 0.541) and Basque (F1 = 0.550); Llama-3-8B performs best on Spanish (F1 = 0.529); Aya performs worse on all languages. This suggests that Qwen may have stronger English and Basque knowledge from pretraining, while Llama may have stronger Spanish and Bangla knowledge, and both Qwen and Llama likely have better multilingual and/or biomedical domain knowledge than Aya. These results reveal the im-

	Age	Symptom	Medicine	Health_Condition	Specialist	Medical_Procedure
json_zero	0.4211	0.0674	0.4000	0.4000	0.5143	0.1333
json_few	0.2000	0.1978	0.4898	0.4000	0.6207	0.1667
json_few_qa	**0.4706**	0.2151	**0.6000**	**0.5333**	0.6000	**0.5455**
json_few_trans	0.0099	0.1957	0.5600	0.3333	**0.6250**	0.2222
json_few_expl	0.2353	**0.2526**	**0.6000**	0.5185	0.5000	0.3636

Table 3: Entity-type–wise F1 scores for Bangla biomedical NER under five prompting strategies. Bold indicates the best F1 per entity type.

	Disorder	Patient	H-Professional
json_zero	0.2153	0.3678	0.1401
json_few	0.3137	0.4413	0.1833
json_few_qa	**0.4268**	**0.5151**	**0.5576**
json_few_trans	0.3795	0.4100	0.3195
json_few_expl	0.4108	0.4093	0.3278

Table 4: Entity-type-wise F1 scores for Basque biomedical NER under five prompting strategies. Bold indicates the best F1 per entity type.

Model	Bangla	Basque	Spanish	English
Llama-3-8B	**0.494**	0.503	**0.529**	0.457
Qwen3-8B	0.399	**0.550**	0.514	**0.541**
Aya	0.229	0.282	0.349	0.318

Table 5: Prompting Strategy Comparison for Multilingual Biomedical NER Using Meta-Llama-3-8B and question-style prompting (json_few_qa).

portance of LLM selection when working with low-resource languages.

5 Error Analysis

5.1 Error Categories

We inspected the errors of the best performing models in Bangla and Basque. Below are the main failure modes, with the operational definitions used in our evaluation. Examples of errors are given in Appendix D and Appendix E.

Hallucination The gold annotation contains no entities, but the model outputs one or more spans.

All-Missed The gold contains entities, but the model returns an empty JSON. This is the most severe recall failure.

Missed Entities The model extracts some entities but misses others, without adding extra entities.

Extra Entities The model predicts additional entities that are not in the gold but does not miss gold entities.

Type Confusion The predicted entity span matches the gold text, but the assigned type is wrong.

Mixed Errors The model both adds and misses entities. Extra-dominant cases indicate over-extraction, while missed-dominant cases indicate under-extraction.

Boundary Mismatch The model captures the right concept but with different boundaries, becoming both a false positive and a false negative under exact-match scoring.

5.2 Error Analysis Explanation

Figure 7 shows clear differences across prompting methods for Bangla biomedical NER errors. Although we have seen that QA-style prompts perform better on average, here we see that translation-based prompting reduces over-generation: it has the lowest HALLUCINATION (2.6%) and comparatively low BOUNDARY MISMATCH (13%). For span-level precision, question-style prompting yields the largest BOUNDARY MISMATCH (29%), suggesting it identifies the right concept but struggles to copy the exact multiword span. Thus, question-style prompting helps the model extract more entities, while translation-based prompting helps reduce hallucination, extra entities, and boundary mismatch errors.

For Basque (Figure 8), across all prompts, HALLUCINATION dominates the errors (about 38–52%), indicating that over-generation is the primary failure mode in Basque. However, the prompts shift the secondary errors: simple statement-based prompting (*json_few*) noticeably reduces HALLUCINATION (37.5%) compared to the other methods, but it increases ALL-MISSED (14.8%), suggesting a more conservative behavior that sometimes fails to extract anything. In contrast, question-style prompting extracts more aggressively and keeps ALL-MISSED low (5.8%), but this comes

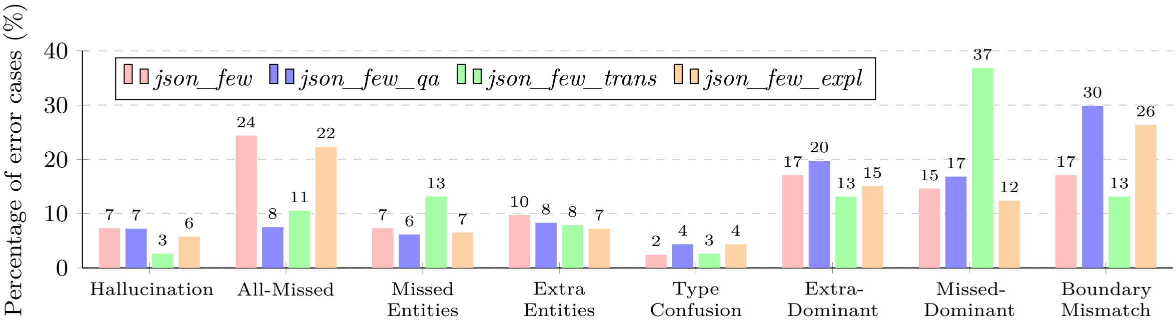

Figure 7: Error type distribution for LLM-extracted entities for Bangla.

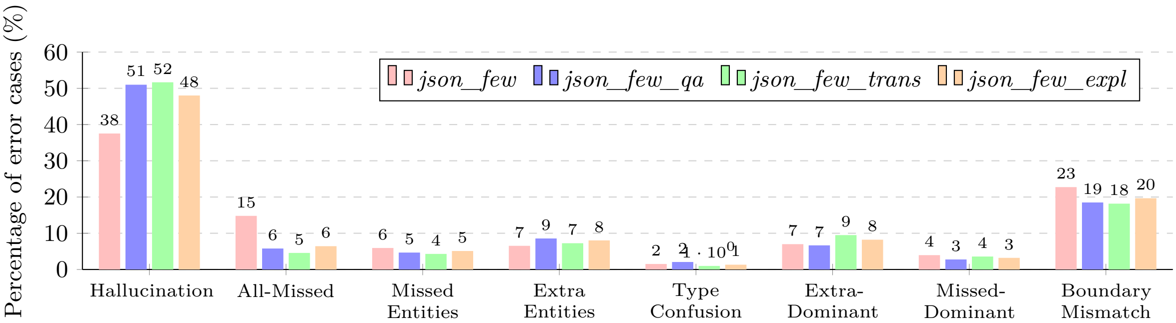

Figure 8: Error type distribution for LLM-extracted entities for Basque.

with higher Extra Entities (8.6%) and substantial Boundary Mismatch (18.5%).

6 Conclusion and Future Work

We systematically evaluated prompt design choices and output schemas for multilingual biomedical NER, focusing on the low-resource, morphologically rich languages Bangla and Basque, while comparing against the high-resource languages Spanish and English. Overall, question-style prompting is the most consistent way to improve exact-match entity-level F1 across languages. Our error analysis also shows clear trade-offs: question-style prompts usually help recall (fewer empty outputs), while translation-based prompting can make the model more conservative and reduce hallucinations and boundary errors in some settings. Together, these results provide practical guidance for building more robust biomedical NER pipelines in low-resource languages without fine-tuning.

We also find that BIO-formatted output is less suitable for LLMs than text-span based outputs. This contrasts with available datasets: not only biomedical NER, but nearly all Bangla NER datasets are annotated in the BIO format (e.g., Islam et al., 2022; Sazzed, 2022; Khan et al., 2023; Mahtab et al., 2025). BIO tagging requires an LLM to assign a correct label to every token and maintain strict consistency. In practice, this is fragile: the model must keep the exact token order and token count, and even one extra or missing token breaks alignment. This becomes worse when tokenization is difficult, as in morphologically rich languages like Bangla and Basque, where suffixes and punctuation frequently change token boundaries. In contrast, text-span-based extraction simply asks the model to list the entity strings in the sentence, which matches how LLMs naturally respond and is easier to parse and store as JSON.

In the future, we plan to test richer question prompts that include short label definitions and examples (not only WH-questions) to see whether they further reduce boundary mismatches and type confusion. We also want to study better post-processing and evaluation for near-miss spans and explore lightweight adaptation methods that choose the best prompt per language and entity type. Finally, expanding to more low-resource languages and more medical datasets will help confirm how general these findings are.

Limitations

A key limitation of this study is that we evaluate only three open LLMs around 8B parameters and two datasets, Bangla HealthNER and Basque E3C, so further work is needed to determine whether the findings generalize to other model sizes, closed models, or biomedical text styles. We also do not cover additional low-resource languages because validating prompts, examples, and error cases reliably requires language expert support. In addition, we did not evaluate paid or proprietary GPT models, and our strict exact match scoring can penalize near-correct span boundaries in morphologically rich languages.

References

Sayantan Adak, Pauras Mangesh Meher, Paramita Das, and Animesh Mukherjee. 2025. REVerSum: A multi-staged retrieval-augmented generation method to enhance Wikipedia tail biographies through personal narratives. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 732–750, Abu Dhabi, UAE. Association for Computational Linguistics.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Dhananjay Ashok and Zachary C. Lipton. 2023. PromptNER: Prompting For Named Entity Recognition. *arXiv preprint*. ArXiv:2305.15444 [cs].

Reza Averly and Xia Ning. 2025. Entity decomposition with filtering: A zero-shot clinical named entity recognition framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2935–2951, Albuquerque, New Mexico. Association for Computational Linguistics.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266, Albuquerque, New Mexico. Association for Computational Linguistics.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. BanglaByT5: Byte-level modelling for Bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5551–5560, Suzhou, China. Association for Computational Linguistics.

Jiawei Chen, Qing Liu, Hongyu Lin, Xianpei Han, and Le Sun. 2022. Few-shot named entity recognition with self-describing networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5711–5722, Dublin, Ireland. Association for Computational Linguistics.

Xiang Cheng, Chengyan Pan, Minjun Zhao, Deyang Li, Fangchao Liu, Xinyu Zhang, Xiao Zhang, and Yong Liu. 2025. Revisiting chain-of-thought prompting: Zero-shot can be stronger than few-shot. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13533–13554, Suzhou, China. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, 49(3):613–641.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818, Torino, Italia. ELRA and ICCL.

Tanvir Islam, Sakila Mahbin Zinat, Shamima Sukhi, Zakir Hossain Zamil, Aynur Nahar, and M. F. Mridha. 2022. An attention-based medical ner in the bengali language. In G. Mathur and 1 others, editors, *Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications*, Algorithms for Intelligent Systems, pages 131–140. Springer Nature Singapore.

Alvi Khan, Fida Kamal, Nuzhat Nower, Tasnim Ahmed, Sabbir Ahmed, and Tareque Chowdhury. 2023. NERvous about my health: Constructing a Bengali medical named entity recognition dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5768–5774, Singapore. Association for Computational Linguistics.

Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Sunayana Sitaram, Kalika Bali, Tanuja Ganu, and Akshay Nambi. 2025. Bridging the language gap: Dynamic learning strategies for improving multilingual performance in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9209–9223, Abu Dhabi, UAE. Association for Computational Linguistics.

Mingchen Li, Yang Ye, Jeremy Yeung, Huixue Zhou, Huaiyuan Chu, and Rui Zhang. 2023a. W-procer: Weighted Prototypical Contrastive Learning for Medical Few-Shot Named Entity Recognition. *arXiv preprint*. ArXiv:2305.18624 [cs].

Mingchen Li and Rui Zhang. 2024. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023b. Crosslingual retrieval augmented in-context learning for Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 136–151, Singapore. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Guizhen Chen, Xiaobao Wu, Anh Tuan Luu, Chip Hong Chang, and Lidong Bing. 2023. Zero-shot text classification via self-supervised tuning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1743–1761, Toronto, Canada. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France. European Language Resources Association.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The E3C project:collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 190–196, Bologna, Italy. CEUR Workshop Proceedings.

Md. Motahar Mahtab, Faisal Ahamed Khan, Md. Ekramul Islam, Md. Shahad Mahmud Chowdhury, Labib Imam Chowdhury, Sadia Afrin, Hazrat Ali, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2025. BanNERD: A benchmark dataset and context-driven approach for Bangla named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6807–6828, Albuquerque, New Mexico. Association for Computational Linguistics.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022. Simple yet powerful: An overlooked architecture for nested named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Salim Sazzed. 2022. BanglaBioMed: A biomedical named-entity annotated corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Sheikh Shafayat, H M Quamran Hasan, Minhajur Rahman Chowdhury Mahim, Rifki Afina Putri, James Thorne, and Alice Oh. 2024. BEnQA: A question answering benchmark for Bengali and English. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177, Bangkok, Thailand. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

IV Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Roberto Zanoli, Alberto Lavelli, Daniel Verdi do Amarante, and Daniele Toti. 2024a. Assessment of the E3C corpus for the recognition of disorders in clinical texts. *Natural Language Engineering*, 30(4):851–869.

Roberto Zanoli, Alberto Lavelli, Daniel Verdi do Amarante, and Daniele Toti. 2024b. Assessment of the E3C corpus for the recognition of disorders in clinical texts. *Natural Language Engineering*, 30:851–869.

Span-level JSON Format

```
{
 "ID": 5,
 "Input": "বেশ কয়েকদিন যাবত
গলায় খুব ব্যাথা সেই সাথে কাশি হয়।
কাশির সাথে ঘন কফ বেরহয় আর
বুকে খুব ব্যাথা হয়। জর ও আছে।",
   "Output": {
  "Symptom": [
   "গলায় খুব ব্যাথা",
   "কাশি",
   "কাশির সাথে ঘন কফ",
   "বুকে খুব ব্যাথা",
   "জর"
     ]
   }
}
```

Token-level BIO Format

```
{
  "ID": 5,
  "TOKEN":[
  "বেশ","কয়েকদিন","যাবত","গলায়","খুব",
  "ব্যাথা","সেই","সাথে","কাশি","হয়","।","কাশির",
  "সাথে","ঘন","কফ","বেরহয়", "আর",
  "বুকে","খুব","ব্যাথা",
  "হয়","।","জর","ও","আছে","।”
]
  "NER_TAG”: [
   "O","O","O","B-Symptom","I-Symptom",
   "I-Symptom", "O","O","B-Symptom","O",
   "O","B-Symptom", "I-Symptom","I-Symptom",
   "I-Symptom","O","O", "B-Symptom",
   "I-Symptom","I-Symptom","O","O",
   "B-Symptom","O","O","O”
]
 }
```

Figure 9: Comparison of span-level JSON and token-level BIO formats for Bangla biomedical NER.

A BIO Tagging vs Span-Based

The left panel of Figure 9 shows span-based extraction, where entities are directly returned as text spans grouped by type in a JSON structure, while the right panel shows token-level BIO tagging, which requires assigning a label to every token and maintaining strict alignment. This example illustrates how span-based formats provide a simpler and more robust representation for LLM-based NER, especially for morphologically rich languages like Bangla.

B Prompting Behavior and Qualitative Analysis for Bangla

Table 6 illustrates qualitative differences across three prompting strategies for Bangla biomedical NER. In the question-style prompt, the model correctly identifies most entities but

<table>
<thead>
<tr><th>E.g.</th><th>Prompt Type</th><th>Prompt</th><th>Llama3-8B</th><th>Gold Label</th></tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>json_-few_qa</td>
<td>You are a question-answering assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Question: Which entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in this text?

Example 1:
এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না।
Answer: Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির.

Now extract all entities for this text:
ধন্যবাদ আপনার প্রশ্নের জন্য। ফাটিয়ে দিবেন না। ইনফেকশন হয়ে যাবে। আপাতত এলারজি উদ্রেককারী খাবার, ধূলাবালি এড়িয়ে চলুন। এলাট্রল খান। না দেখে সমাধান দেয়া যাচ্ছে না। সরাসরি চর্মরোগ বিশেষজ্ঞ দেখিয়ে পরামর্শ নিতে হবে ধন্যবাদ</td>
<td>Medicine: এলাট্রল Specialist: চর্মরোগ বিশেষজ্ঞ</td>
<td>Medicine: এলাট্রল Health_-Condition: ইনফেকশন Specialist: চর্মরোগ বিশেষজ্ঞ</td>
</tr>
<tr>
<td>2</td>
<td>json_-few_-trans</td>
<td>You are a NER assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Example 1:
এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না।

English translation of the Example: Because of allergy problems, when I go out in the sun my body tingles, and I feel a crawling sensation inside my head. In the morning after waking up I sneeze continuously. I have never taken any allergy medicine regularly. In this case, what can I do? Because of this allergy, I cannot do many daily tasks. Now findout the entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in this text.

Answer: Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির.

Now extract all entities for this text:
আপনার বাচ্চাকে আদা, মধু, গরম পানি খাওয়ান। সিরাপ এমব্রোক্স খাওয়াতে পারেন আধা চামচ করে তিন বার একজন শিশু বিশেষজ্ঞ কে দেখিয়ে নিন।</td>
<td>Medicine: আদা | গরম পানি | মধু | সিরাপ এমব্রোক্স ; Specialist: শিশু বিশেষজ্ঞ</td>
<td>Medicine: সিরাপ এমব্রোক্স ; Specialist: শিশু বিশেষজ্ঞ</td>
</tr>
<tr>
<td>3</td>
<td>json_-few_-expl</td>
<td>You are a NER assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Example 1:
এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না।

Explanation of this Example: The Bangla text clearly describes an allergy problem, which is a Health_Condition entity (“এলার্জির”). It also describes allergy-related symptoms, such as skin discomfort in the sun, crawling sensation in the head, and continuous sneezing (Symptom). It mentions “এলার্জির ওষুধ” only in a generic way, but it still refers to medicine use (Medicine). No age, specialist, or procedure is stated. Now read the Bangla text and check the Answer, then identify which entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in the text.

Answer: Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির.

Example 2:
রোগীর বয়স ৫৫ বছর। তিনি ডায়াবেটিস ও উচ্চ রক্তচাপের রোগী এবং মেটফরমিন ও লোসারটান খাচ্ছেন।
Answer: Age – ৫৫ বছর. Health_Condition – ডায়াবেটিস | উচ্চ রক্তচাপ.

Now extract all entities for this text:
আমার বয়স ১৩ আমার পচন্ড পায়ের হার এর বেথা। আমি কি ঔষধ খাব বলেন স্যার। Calcium BP 500mg কিনছি খাব কি স্যার।</td>
<td>Age: ১৩ বছর ; Symptom: পচন্ড পায়ের হার এর বেথা ; Medicine: Calcium BP 500mg ; Specialist: স্যার</td>
<td>Age: ১৩ ; Symptom: পায়ের হার এর বেথা ; Medicine: Calcium BP 500mg</td>
</tr>
</tbody>
</table>

Table 6: Example prompt templates for Bangla biomedical NER. We show three prompt variants (question-style, translation-based, and explanation-based) with one in-context demonstration and one target Bangla input. For each example, we report the Meta-Llama-3-8B output and the corresponding gold labels for comparison.

<table>
<tr><th>E.g.</th><th>Prompt Type</th><th>Prompt</th><th>Llama3-8B</th><th>Gold Label</th></tr>
<tr><td>1</td><td>json_-few_qa</td><td>You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Question: Which entities (Disorder, Patient, H-Professional) are present in this text?

Example 1:
Text: "-Eskuineko besapea:aurkikuntza patologikorik ez."
Answer: {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []}

Now extract all entities for this text:
Estenosi aortiko larriari bihotz-gutxiegitasuna eta sinkopea gehitu zaizkionez, ebakuntza egiteko baldintzak betetzen ditu, eta Bihotz-hodietako Kirurgiako Zerbitzuarekin harremanetan jarri gara.</td><td>Disorder: bihotz-gutxiegitasuna | estenosi aortiko larria | sinkopea ; H-Professional: Bihotz-hodietako Kirurgiako Zerbitzuarekin</td><td>Disorder: Estenosi aortiko | bihotz-gutxiegitasuna | sinkopea]</td></tr>
<tr><td>2</td><td>json_-few_-trans</td><td>You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Example 1:
Text: "-Eskuineko besapea:aurkikuntza patologikorik ez."
English translation of the text: "-Right armpit: no pathological findings."
Answer: {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []}

Now extract all entities for this text:
Txankro sifilitikoa/Sifilis primarioa.</td><td>Disorder: Sifilis primarioa | Txankro sifilitikoa</td><td>Disorder: Sifilis | Txankro sifilitikoa</td></tr>
<tr><td>3</td><td>json_-few_-expl</td><td>You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON.

Example 1:
Text: "-Eskuineko besapea:aurkikuntza patologikorik ez."
Description of text: A clinical examination of a right armpit with no health disorders or pathological findings identified. You have to find out which entities (Disorder, Patient, H-Professional) are present in the text.
Answer: {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []}

Now extract all entities for this text:
Edemarik ez; ez dago bena sakonetako tronbosi-zeinurik.</td><td>Disorder: tronbosi-zeinurik</td><td>Disorder: Edemarik | bena sakonetako tronbosi-zeinurik</td></tr>
</table>

Table 7: Example prompt templates for Basque biomedical NER. We show three prompt variants (question-style, translation-based, and explanation-based) with one in-context demonstration and one target Basque input. For each example, we report the Meta-Llama-3-8B output and the corresponding gold labels for comparison.

misses the HEALTH_CONDITION (ইনফেকশন), showing that QA prompting can still suffer from recall gaps. In contrast, translation-based prompting leads to over-generation: the model incorrectly treats common home remedies such as গরম পানি (hot water) and মধু as MEDICINE, likely due to reasoning in English and broader semantic interpretation. Explanation-based prompting improves recall but introduces type confusion, as the model incorrectly labels the polite address স্যার ("Sir") as a SPECIALIST. These examples highlight that while extended prompting strategies can help extraction, they also introduce distinct error patterns depending on how semantic cues are framed.

C Prompting Behavior and Qualitative Analysis for Basque

Table 7 presents qualitative differences in model behavior across three prompting strategies for Basque biomedical NER. Under the question-style prompt, the model successfully identifies major disorders but exhibits boundary and normalization issues, such as over-specifying severity or partially mismatching gold entity forms. Translation-based prompting generally improves coverage by leveraging English semantics, but it can introduce label drift, as closely related clinical concepts are normalized differently from the gold labels. Explanation-based prompting further increases recall by encouraging semantic inference, yet it also leads to over-generation, with the model extracting implicit or negated find-

<table>
<tr><th>Error Type</th><th>Input Example</th><th>Gold Output</th><th>Pred Output</th></tr>
<tr><td>Hallucinations</td><td>ধন্যবাদ আপনাকে প্রশ্নের জন্য। জ্বর ৯৯ হলে ওষুধ খাওয়াবেন। নয়ত দরকার নেই। ধন্যবাদ</td><td>Empty</td><td>Symptom: জ্বর</td></tr>
<tr><td>All-Missed</td><td>আমি। আমার বয়স 27 বছর। সাম্প্রতিক সময়ে আমি আমার শরীরে অস্বস্...ান্ত হয়ে গেছি, আগের থেকে শরীরের ওজন বৃদ্ধি পেয়েছে, পেটে প্র...</td><td>Age: 27 বছর ; Symptom: একাধারে মাটিতে বসে থাকলে মাজার ...| শরীরের ওজন বৃদ্ধি | হাটলেই মনে হয় ক্লান্ত হয়ে</td><td>Empty</td></tr>
<tr><td>Missed Entities</td><td>total cholesterol 200 mg / dl serum triglycerides 553 mg / dl...how to lower triglycerides? am i prone to heart disease?</td><td>Symptom: HDL cholesterol 25 mg | LDL cholesterol 93 mg | cholesterol...| serum triglycerides 553 mg ; Health_Condition: heart disease</td><td>Health_Condition: heart disease</td></tr>
<tr><td>Extra Entities</td><td>আপনাকে ধন্যবাদ প্রশ্ন করার জন্য। আপনি গরম পানির ধোয়া নাকে নিয়...রাতে একটা করে খান। একজন মেডিসিন বিশেষজ্ঞ পরামর্শ নিন। ধন্যবাদ</td><td>Medicine: এলাট্রল ; Specialist: মেডিসিন বিশেষজ্ঞ</td><td>Medicine: এলাট্রল | গরম পানির ধোয়া ; Specialist: মেডিসিন বিশেষজ্ঞ</td></tr>
<tr><td>Type Confusion</td><td>আমার স্ত্রী বয়স ২৪ ও ১৪ দিনের প্রেগন্যান্ট অবস্থা MM Kit খায় ...আল্টাস্রোফ্রাগি করাতে চাই, দয়া করে আমাকে পরিক্ষার নাম টা লিখে দিবেন</td><td>Age: ২৪ ; Symptom: bleeding আনে বেশি যাচ্ছে ; Medicine: MM Kit ...; Health_Condition: মাসিক | ১৪ দিনের প্রেগন্যান্ট</td><td>Age: ১৪ দিন | ২৪ ; Symptom: bleeding ; Medicine: MM Kit ; Health_Condition: প্রেগন্যান্ট অবস্থা ; Medical_Procedure: আল্টাস্রোফ্রাগি</td></tr>
<tr><td>Mixed Errors</td><td>Thank you for your question. Your serum Triglyceride is sligh...with your reports. Avoid fatty food, carbohydrate. Thank you.</td><td>Symptom: serum Triglyceride is slightly raised</td><td>Medicine: drug ; Specialist: doctor</td></tr>
<tr><td>Boundary Mismatch</td><td>বেশ কয়েকদিন যাবত গলায় খুব ব্যাথা সেই সাথে কাশি হয়। কাশির সাথে ঘন কফ বেরহয় আর বুকে খুব ব্যাথা হয়। জর ও আছে।</td><td>Symptom: কাশি | কাশির সাথে ঘন কফ | গলায় খুব ব্যাথা | জর | বুকে খুব ব্যাথা</td><td>Symptom: কাশি | খুব ব্যাথা | ঘন কফ | জর | বুকে খুব ব্যাথা</td></tr>
</table>

Table 8: Representative QA-prompt error examples for Bangla biomedical NER (Llama3.1-8B). Each row shows one real instance of a major error category under exact-match span scoring.

<table>
<tr><th>Error Type</th><th>Input Example</th><th>Gold Output</th><th>Pred Output</th></tr>
<tr><td>Hallucinations</td><td>Burua eta lepoa.</td><td>Empty</td><td>Disorder: Burua eta lepoa</td></tr>
<tr><td>All-Missed</td><td>Ez dago aurkikuntza patologikorik.</td><td>Disorder: aurkikuntza</td><td>Empty</td></tr>
<tr><td>Missed Entities</td><td>Estenosi aortiko larria eta bihotz-gutxiegitasuna.</td><td>Disorder: Estenosi aortiko larria | bihotz-gutxiegitasuna</td><td>Disorder: bihotz-gutxiegitasuna</td></tr>
<tr><td>Extra Entities</td><td>Sabela.</td><td>Empty</td><td>Patient: Sabela</td></tr>
<tr><td>Type Confusion</td><td>Biguna eta zanpagarria.</td><td>Empty</td><td>Disorder: Biguna | Disorder: zanpagarria</td></tr>
<tr><td>Mixed Errors</td><td>Biriketako murmurio normala.</td><td>Empty</td><td>Disorder: Biriketako murmurio</td></tr>
<tr><td>Boundary Mismatch</td><td>Ez dago aurkikuntza patologikorik.</td><td>Disorder: aurkikuntza patologikorik</td><td>Disorder: aurkikuntza</td></tr>
</table>

Table 9: Representative QA-prompt error examples for Basque biomedical NER (Meta-Llama-3-8B). Each row illustrates a distinct error category under exact-match span evaluation.

ings (e.g., absence of edema or thrombosis) as entities. Overall, these examples demonstrate that richer prompting strategies can enhance extraction in Basque but also introduce systematic errors related to semantic inference, normalization, and negation handling.

D Bangla Errors (QA Prompt)

Table 8 presents representative Bangla error cases for the QA-style prompt. A recurring pattern is that the model often identifies the correct medical concept but fails to reproduce the exact mention boundaries, especially for multiword symptoms and colloquial expressions, resulting in **Boundary Mismatch**. We also observe **Hallucinations** when generic medical advice or common words are treated as symptoms/conditions even when the gold annotation is empty. In contrast, **All-Missed** cases reflect severe recall failures where the model returns an empty JSON despite clear entity cues (e.g., explicit age or specialist mentions). Finally, **Type Confusion** commonly arises between semantically adjacent categories (e.g., tests/procedures vs. conditions, or informal medication names vs. remedies), indicating that the model recognizes the span but struggles with fine-grained label assignment.

E Basque Errors (QA Prompt)

Table 9 shows representative Basque error cases for the same QA-style prompt. Across prompts, the dominant failure mode is **Hallucinations**, suggesting that the model is prone to over-generation in Basque medical text: short phrases and anatomical references are frequently over-labeled as medical entities even when the gold annotation contains none. We also see **Boundary Mismatch** in cases where the model selects a shorter head noun rather than the full descriptive mention, which is penalized under exact-match evaluation. Compared with Bangla, Basque exhibits more frequent over-extraction and fewer purely recall-only failures, consistent with the overall error distribution where FP-heavy errors occupy a larger fraction of error-only cases.