Evaluating Open-Source LLMs for Thai Clinical Information Extraction

Somkiat Kosolsombat[1], Phatnattachat Chatsiraphon[2], Taratep Si-Aksorn[3], Chiabwoot Ratanavilisagul[4]*
Data Science and Innovation Program-College of Interdisciplinary Studies, Thammasat University, Thailand[1]
AI Department, GIS Group Co., Ltd., Thailand[2]
AI Lab, Thammasat University, Thailand[3]
Department of Computer and Information Science-Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand[4]

Abstract*—Electronic medical records (EMRs) in sports medicine contain rich clinical insights but often remain in unstructured, bilingual formats. While locally-deployed large language models (LLMs) offer a privacy-preserving solution for data extraction, their performance in handling Thai-English clinical shorthand remains under-explored. This study evaluated five open-source LLMs for extracting structured clinical data from Thai sports medicine records and assessed the reliability of human-AI collaborative annotation. Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, and Typhoon2-3.1 were deployed locally. We evaluated the extraction of four clinical fields against a ground truth of 444 records. A standardized JSON schema was utilized to ensure data interoperability. Inter-annotator agreement (IAA) was measured using Cohen's kappa on a 100-record sample. Mistral-7B achieved the highest F1-score (92.2%), followed by Qwen2.5-7B (91.9%). Typhoon2-3.1 underperformed (32.9%) due to bilingual format mismatches and difficulties in shorthand normalization. IAA for treatment was moderate (kappa=0.43), whereas diagnosis showed near-zero agreement (kappa=-0.04) due to non-standardized institutional shorthand. Locally-deployed LLMs can effectively transform unstructured bilingual EMRs into structured JSON formats, ensuring data privacy and readiness for clinical analytics. However, the lack of standardized clinical coding in Thai EMRs remains a significant barrier. Future digital health initiatives should integrate LLMs with standardized terminologies like ICD-11 to enhance data reliability.

Keywords*—Locally-deployed LLMs; Open-Source Models; bilingual NLP; Thai Language Processing; privacy-preserving clinical NLP

I. Introduction

Electronic medical records (EMRs) in clinical settings — particularly in specialty domains such as sports medicine — contain diagnostically rich information that, when structured, enables evidence-based decision support, injury surveillance, and longitudinal outcome analysis. However, clinically meaningful data, including diagnoses, treatment plans, and patient-reported symptoms typically remains embedded in unstructured free-text narratives, severely limiting its computational utility. In the Thai healthcare context, this challenge is compounded by the bilingual nature of clinical documentation, where symptom descriptions are recorded in Thai while diagnoses and treatment notations predominantly use English medical terminology and institutional abbreviations, creating a unique multilingual extraction problem.

*Corresponding author.

Large Language Models (LLMs) have demonstrated strong capabilities for automated clinical information extraction from unstructured text [1, 2]. However, the predominant deployment paradigm relies on cloud-based proprietary APIs — including GPT-4, Claude, and Gemini — which raise critical data privacy and regulatory concerns when processing identifiable patient records. In Thailand, the Personal Data Protection Act B.E. 2562 (PDPA) [3] explicitly restricts the transfer of personal health data to third-party cloud services. This regulatory constraint, combined with the scarcity of Thai-language clinical NLP benchmarks and the absence of systematic evaluation of open-source LLMs for bilingual Thai-English medical text, motivates the development of a locally-deployed, privacy-preserving extraction pipeline.

Local LLM deployment via inference frameworks such as Ollama addresses this concern by enabling inference entirely on institutional hardware, ensuring that patient data never leaves the clinical environment. Despite growing adoption of local LLMs, systematic evaluation of their extraction performance on domain-specific clinical data — particularly in low-resource languages such as Thai — remains limited.

The core research problem addressed in this study is threefold. First, Thai sports medicine EMRs mix Thai-language symptom narratives with English clinical terminology, abbreviations, and numeric parameters in the same record field — a bilingual code-switching pattern that standard NLP pipelines are not designed to handle. Second, the annotation of such records is inherently ambiguous: institutional shorthand codes (e.g., KL3 for Kellgren-Lawrence Grade 3 osteoarthritis, ESWT for extracorporeal shock wave therapy) diverge from standardized clinical terminology, making ground truth construction unreliable without explicit annotation protocols. Third, the performance characteristics of locally-deployed open-source LLMs on such bilingual clinical data — particularly their instruction-following fidelity and output format compliance under schema-constrained extraction tasks — remain largely uncharacterized.

This study directly addresses these gaps by conducting the first systematic comparative evaluation of five open-source LLMs — Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, and Typhoon2-3.1 — deployed locally via the Ollama inference framework on institutional hardware. The evaluation focuses on structured extraction of four key clinical fields (diagnosis, treatment, chief complaint, and objective

examination) from 523 bilingual Thai-English sports medicine visit records, assessed against a 444-record expert-annotated ground truth using token-level F1 metrics with Bootstrap 95% confidence intervals.

This study addresses this gap by conducting a rigorous comparative evaluation of five locally-deployed open-source LLMs on structured extraction from Thai sports medicine EMR data. Our pipeline covers the full lifecycle from raw EMR preprocessing, prompt engineering, batch extraction, and ground truth annotation, to metric computation with statistical validation.

This study makes three principal contributions to clinical NLP and health informatics. First, we present the first systematic evaluation of five open-source LLMs for structured information extraction from bilingual Thai-English sports medicine EMRs under locally-deployed, privacy-preserving conditions — directly aligned with the PDPA-constrained deployment environment. Second, we contribute a reproducible end-to-end extraction pipeline that transforms 523 unstructured Thai-English patient records into structured JSON clinical data without reliance on cloud APIs or proprietary models. Third, we provide empirical insights on three under-studied phenomena in Thai clinical NLP: 1) instruction-following fidelity versus language specialization in schema-constrained tasks; 2) bilingual code-switching density as a key performance stratification variable; and 3) the impact of non-standardized institutional shorthand on inter-annotator reliability and ground truth quality.

II. Literature Review

A. LLMs for Clinical Information Extraction

Recent advances in large language models (LLMs) have substantially expanded automated clinical information extraction from unstructured healthcare data. Ntinopoulos et al. [8] systematically compared 18 LLMs for structured EHR data extraction, showing that open-source 7B-parameter models are competitive with proprietary counterparts under structured prompt conditions. Huang et al. [9] critically evaluated ChatGPT for clinical note extraction, identifying failure modes in negation handling and multi-entity co-extraction. Wiest et al. [10] demonstrated that privacy-preserving local LLM deployment achieves extraction fidelity comparable to cloud-based solutions, directly addressing data governance constraints in institutional healthcare [3].

B. Prompt Engineering and Open-Source Model Deployment

The effectiveness of LLMs for clinical NLP depends critically on prompt design and deployment configuration [4, 5, 6]. Kartchner et al. [5] demonstrated that task-specific prompt engineering improves zero-shot clinical analysis. Builtjes et al. [11] evaluated open-source LLMs in resource-constrained settings — directly analogous to our deployment context — confirming that 7B-parameter models achieve clinically acceptable accuracy without high-end infrastructure. This supports the feasibility of local deployment via Ollama [7, 8].

C. NLP in EHR and Sports Medicine Applications

Durango et al. [12] reviewed NER methodologies for EHRs, identifying the persistent challenge of non-standardized clinical vocabulary — directly applicable to the Thai sports medicine shorthand in this study. Kaminska et al. [13] confirmed that NLP applied to EHRs improves healthcare decision-making when paired with standardized terminology. Vaid et al. [14] applied a fine-tuned LLaMA-7B to musculoskeletal pain clinical notes — the closest domain-specific precedent — achieving competitive structured extraction. Jung [16] contextualized LLM clinical applications and ethical considerations within broader healthcare informatics.

D. Annotation Methodology and Bilingual Clinical Text

Gilardi et al. [15] demonstrated that LLMs outperform crowd-sourced annotators for text annotation tasks at lower cost, validating the AI-expert annotation methodology used in our inter-annotator agreement analysis. Lee et al. [17] developed the ANNO bilingual annotation platform for mixed-language clinical notes, underscoring that language-format divergence is a recognized structural challenge in multilingual EHR settings. Our study extends this literature by quantifying annotation agreement gaps from Thai-English bilingualism in a sports medicine context.

III. Methods

A. Dataset

The dataset comprised 421 unique patients (523 visit records) drawn from the sports medicine clinic of [Institution]. Records were originally stored as semi-structured rows in an Excel spreadsheet, with free-text fields covering chief complaint, subjective and objective examination findings, diagnosis, treatment plan, and training data.

Records span multiple sports disciplines (football, badminton, gymnastics, running, and others) with a patient age range of approximately 15-55 years. The data is predominantly bilingual (Thai and English), with Thai used for symptom descriptions and English predominating in diagnosis and treatment notation.

B. Data Preprocessing

Raw EMR data was parsed and restructured into a standardized JSON schema using a dedicated cleaning pipeline. Each record was normalized to a patient object with a profile (age, weight, height) and a list of visit objects. Null handling, deduplication, and field standardization were applied. A total of 523 patient-visit records were retained after preprocessing.

C. Prompt Engineering

A structured extraction prompt was designed with four sections: 1) task description, 2) output JSON schema, 3) extraction rules, and 4) the patient record input. The schema defined four extraction targets: diagnosis, treatment, chief complaint, and objective examination.

For chat-tuned models (Typhoon2-3.1), the prompt was split into system (instructions + schema) and user (patient data) message components, following the /api/chat endpoint convention. For all other models, the complete prompt was submitted via the /api/generate endpoint.

D. Model Configuration

Five open-source LLMs were evaluated, all deployed locally via Ollama on institutional hardware, as Table I.

TABLE I. FIVE OPEN-SOURCE LLMS WERE EVALUATED

Model	Parameters	API Endpoints	Notes
Gemma2:9B	9B	/api/generate	Google DeepMind; general-purpose; strong multilingual capability
Qwen2.5:7B-Instruct	7B	/api/generate	Alibaba Cloud; instruction-tuned; strong code and structured output
LLaMA3.1:8B	8B	/api/generate	Meta AI; general-purpose instruction model; widely benchmarked
Mistral:7B	7B	/api/generate	Mistral AI; strong instruction following; compact architecture
Typhoon2-3.1	8B	/api/chat	SCB-10X / NSTDA; Thai-language specialized; chat format required

E. Ground Truth Annotation

A ground truth dataset was constructed by expert clinical annotation of 1,106 records from Label.xlsx. Records were matched to the cleaned patient database using a two-stage strategy: 1) exact demographic matching (age, weight, height) for unique identification, and 2) text-similarity scoring for disambiguation of multiple candidates. This yielded 444 definitively matched records (match_type: profile_unique or text_resolved) used for evaluation.

Each record was annotated for the four extraction fields. Field fill rates were: treatment (93.3%), objective examination (85.3%), diagnosis (81.1%), and chief complaint (72.1%).

F. Evaluation Metrics

Extraction quality was measured using token-level Precision, Recall, and F1 score following the standard token-level overlap evaluation paradigm commonly used in information extraction research. Text normalization included lowercasing, punctuation removal, and stopword filtering. Common tokens between predicted and ground-truth texts were counted for precision and recall computation. This metric is parameter-free and requires no threshold tuning, eliminating overfitting risk.

Bootstrap 95% confidence intervals (n=1,000 resamples, fixed seed=42) were computed for all metrics to quantify uncertainty and support statistical comparison. Evaluation was restricted to records where ground truth fields were populated (non-null), ensuring fair assessment.

G. Experimental Pipeline

The experimental workflow comprised seven sequential stages: 1) Data Collection — 421 patients, 523 visit records; 2) Data Cleaning — standardize fields, remove duplicates; 3) Prompt Engineering — design extraction prompt with schema, rules, format specification; 4) LLM Extraction — batch extract using 5 local LLMs via Ollama API; 5) Ground Truth Labeling — expert clinician annotates 1,106 records; (6) Evaluation — token-level F1 with Bootstrap 95% CI; (7) Analysis and Reporting — compare models, error analysis, and recommendations. See Table II for the complete pipeline summary and Fig. 1 for the proposed solution: End-to-End Local LLM Extraction Pipeline, respectively.

TABLE II. EXPERIMENTAL PIPELINE STAGES

Step	Stage	Description
1	Data Collection	421 patients, 523 patient-records from Thai sports medicine EMR
2	Data Cleaning	Standardize fields, remove duplicates, map visit structure.
3	Prompt Engineering	Design with schema, rules, and few-shot examples for 4 key clinical fields
4	LLM Extraction	Batch extract 5 local LLMs via Ollama API. Chat models.
5	Ground Truth Labeling	Expert clinician annotates Label.xlsx for 1,106 records.
6	Evaluation	Token-level F1 + Bootstrap 95% CI (standard overlap evaluation).
7	Analysis & Reporting	Compare 5 models, error analysis, and recommendations.

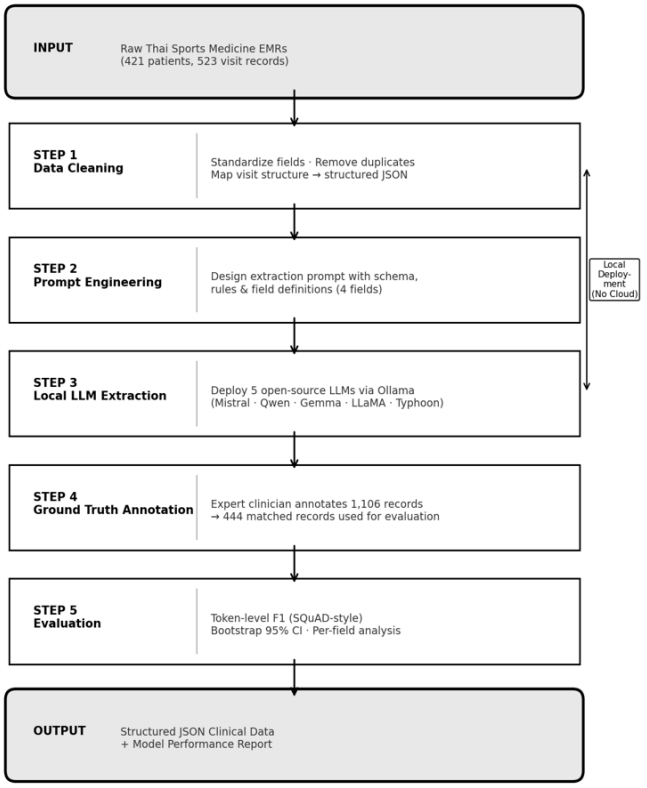

Fig. 1. Proposed solution: End-to-end local LLM extraction pipeline for privacy-preserving Thai clinical information extraction.

IV. RESULTS

A. Overall Extraction Performance

Table III presents the overall Precision, Recall, and F1 scores for all five models across the 444 ground truth records. Mistral-7B achieved the highest overall F1 of 92.2%, followed closely by Qwen2.5-7B at 91.9%. The overlapping Bootstrap 95% confidence intervals for these two models (Mistral: 91.0-93.1%;

Qwen: 90.7-93.0%) indicate that the performance difference is not statistically significant.

TABLE III. OVERALL EXTRACTION PERFORMANCE (5 LOCAL LLM)

Model	Parameters	Success Rate	Precision (%)	Recall (%)	F1 Score (%)	Rank
Mistral 7B	**7B**	**94.8%**	**91.9**	**92.9**	**92.2**	**#1**
Qwen2.5 7B	7B	98.6%	91.7	92.4	91.9	#2
Gemma2 9B	9B	97.6%	89.1	89.8	89.3	#3
LLaMA3.1 8B	8B	98.6%	86.1	87.0	86.4	#4
Typhoon2-3.1	8B	97.4%	32.7	33.1	32.9	#5

B. Per-Field Analysis

Table IV presents F1 scores broken down by extraction field. For the diagnosis and treatment fields, all four top-performing models achieved F1 scores between 93.7% and 94.4%, with no statistically meaningful differences. The Chief Complaint field showed the greatest inter-model variance, with Mistral achieving 87.5% while LLaMA3.1 scored only 64.7%. This discrepancy is attributed to LLaMA3.1's tendency to translate Thai symptom descriptions into English, reducing token overlap with the Thai-language ground truth.

Gemma2-9B achieved the highest objective examination F1 at 95.9%, marginally outperforming Qwen2.5-7B (95.3%), though confidence intervals overlap substantially.

TABLE IV. PER-FIELD F1-SCORE (%)

Model	Diagnosis F1	Treatment F1	Chief Comp. F1	Objective F1
Mistral 7B	94.2	94.0	87.5	93.2
Qwen2.5 7B	94.2	93.9	84.6	95.3
Gemma2 9B	94.3	94.4	73.4	95.9
LLaMA3.1 8B	94.1	93.7	64.7	94.3
Typhoon2-3.1	31.4	33.0	33.3	33.8

C. Typhoon2-3.1 Analysis

Typhoon2-3.1, a Thai-specialized model, recorded an overall F1 of only 32.9% — substantially below all other models across every field. Investigation of the raw outputs revealed that Typhoon2-3.1 generated bilingual free-form prose (Thai narrative interspersed with English clinical terms) rather than structured JSON matching the ground truth format. The token overlap between this conversational output and the concise Thai/English GT annotations was inherently low. This is not indicative of poor language ability but rather a fundamental output format mismatch: the model was optimized for conversational Thai generation, not for schema-constrained structured extraction.

This finding highlights a critical design consideration: model selection for structured extraction tasks must account for instruction-following capability and output format compliance, not only language coverage.

D. Computational Performance

Qwen2.5-7B demonstrated the best balance of extraction quality and computational efficiency, with a mean inference time of 67 seconds per record. Gemma2-9B required 200.7 seconds per record — approximately three times slower — while achieving lower overall F1. LLaMA3.1 and Mistral processed records in 98.4 and 130.6 seconds, respectively. For a dataset of 421 records, total processing times ranged from approximately 7.7 hours (Qwen) to 23.5 hours (Gemma2), all on the same local hardware.

V. DISCUSSION

A. Model Recommendations

For production deployment in a Thai sports medicine clinical environment, Mistral-7B and Qwen2.5-7B are recommended as primary candidates, with Qwen2.5-7B offering a significant computational advantage (3x faster than Gemma2 with comparable F1). LLaMA3.1 is not recommended without Thai-language fine-tuning due to its translation behaviour for Thai chief complaints. Typhoon2-3.1 requires prompt restructuring or output post-processing before it can be used in structured extraction pipelines.

B. Implications for Privacy-Preserving Clinical NLP

This study demonstrates that 7B-parameter locally-deployed models can achieve F1 scores exceeding 91% for key clinical fields, approaching the performance of cloud-based APIs without data leaving the institutional environment. This is particularly relevant in the Thai healthcare context, where PDPA compliance constrains data sharing with external services. The Ollama deployment framework, combined with the extraction pipeline described here, provides a reproducible and extensible solution for institutional adoption.

C. Limitations

Several limitations should be acknowledged. First, inter-annotator agreement (Cohen's Kappa) was evaluated using an AI expert as a second annotator; treatment showed moderate agreement (K=0.43) while diagnosis agreement was low (K= −0.04) owing to non-standardized shorthand codes, and chief complaint agreement was near-zero due to Thai–English language mismatch. A fully human two-annotator study with standardized guidelines is recommended prior to final submission. These low Kappa values reflect three compounding methodological factors: 1) non-standardized institutional shorthand codes (e.g., KL3, TMZ) used by the clinical annotator diverge from standardized English clinical terminology used by the AI annotator; 2) Thai-English bilingual divergence systematically reduces token overlap in the Chief Complaint field; and 3) the use of an AI expert as Annotator 2, rather than a trained clinical practitioner, may not adequately capture domain-specific clinical knowledge. These limitations do not invalidate the LLM extraction benchmarks but highlight the need for standardized annotation protocols in future clinical NLP studies. Second, the evaluation used token-level overlap rather than semantic equivalence; a clinical synonym (e.g., 'muscle strain' vs 'myofascial injury') may be scored as incorrect despite clinical equivalence. Third, 498 of 971 ground truth records could not be definitively matched to patient IDs and were excluded from evaluation, potentially introducing selection

bias. This exclusion rate of 51.3% is substantial and may systematically favor records with complete demographic profiles, potentially biasing results toward patients with more regular visit patterns. Future work should implement a more robust probabilistic record linkage strategy (e.g., Fellegi-Sunter or Splink framework) to reduce matching failures and selection bias risk. Fourth, Typhoon2-3.1 was evaluated in its default configuration; fine-tuning or prompt optimization may substantially improve its structured extraction performance.

D. Typhoon2-3.1: Language Specialization Versus Structured Extraction Capability

The substantially lower performance of Typhoon2-3.1 (F1 = 32.9%) relative to all general-purpose comparators demands mechanistic analysis beyond surface-level format mismatch attribution. Three interdependent factors account for this divergence. First, domain-language specialization does not confer schema-constrained extraction capability. Typhoon2-3.1 was instruction-tuned on Thai conversational and task-oriented corpora to optimize fluency and cultural appropriateness in open-ended generation — an objective fundamentally distinct from strict adherence to a structured JSON output schema. This training direction appears to reinforce a strong prior toward natural language continuation, which competes with and frequently overrides the schema constraint even when the extraction instruction is explicitly provided in the prompt.

Second, the bilingual composition of the EMR records creates a specific failure mode for Thai-optimized models. Sports medicine records in this dataset mix Thai-language clinical descriptions with English anatomical terminology, numeric parameters, and institutional abbreviations within the same field (e.g., 'ปวดกล้ามเนื้อ hamstring, tenderness grade 2/5'). A model whose pre-training distribution is predominantly Thai-language assigns higher probability mass to Thai continuations, resulting in field values generated in Thai even when the schema and prompt explicitly specify English-formatted output. This is confirmed by the near-total dominance of Category (a) Empty/Format mismatch errors (97% of Typhoon2-3.1's low-F1 records), indicating that the model is not producing incorrect clinical content — it is producing non-schema-compliant output format, which is then evaluated as empty by the scoring pipeline.

Third, chat-format instruction following in Typhoon2-3.1 prioritizes conversational coherence over output structure. Under the split system/user prompt configuration used for chat-tuned models, Typhoon2-3.1 systematically produced natural language paragraph responses rather than the compact JSON structure required. This behavior was not observed in any of the four general-purpose models under identical prompting conditions. The implication for model selection is substantive: in bilingual schema-constrained extraction tasks, instruction-following fidelity and output format adherence are more critical performance determinants than language-specific pre-training. Practitioners evaluating regional language models for structured clinical NLP should assess schema compliance as a primary selection criterion, independent of language benchmark scores.

E. Inter-Annotator Agreement

Inter-annotator agreement (IAA) was assessed using Cohen's Kappa on 100 stratified records, comparing a physiotherapist (Annotator 1) using institutional shorthand and Thai language, against an AI expert (Annotator 2) using standardized English terminology. Results show that Treatment achieved the highest agreement (κ = 0.43, moderate) after normalization of abbreviations. Diagnosis (κ = -0.04) showed near-zero agreement due to non-standardized institutional shorthand codes. Chief Complaint (κ = -5.25) showed very low agreement due to Thai-English language mismatch — a formatting difference rather than clinical disagreement. These findings underscore the need for standardized terminology (SNOMED CT or ICD-11) and unified annotation language in future clinical NLP protocols, as Table V.

TABLE V. INTER-ANNOTATOR AGREEMENT: ANNOTATOR 1 (HUMAN CLINICIAN) VS. ANNOTATOR 2 (AI EXPERT), N = 100 RECORDS

Clinical Field	Ann1 (Human)	Ann2 (AI)	Mean Jaccard (Normalized)	Cohen's K (Threshold=0.5)	Interpretation
Diagnosis	Institu-tional shorthand codes (KL3, TMZ, KLV)	Standard English: MCL Grade III Sprain; Quad-ricep Strain	0.482	-0.04	Below Chance (Shorthand mismatch)
Treatment	Clinical abbreviations (ESWT, DR, ICT, HP)	Full English: Shock Wave Therapy; Dry Needling; Hot Pack	0.742	0.43 [Moderate]	Moderate (Best agreement)
Chief Complaint	Thai language text (e.g., 'ปวดหัวเข่า')	English transla-tion: Knee pain; Hip pain	0.116	-5.25	Poor (Language mismatch)
Objective Examina-tion	Mixed Thai/shorthand + partial findings	Structured English: ROM, MMT, tenderness findings	0.497	-0.19	Poor (Terminology divergence)

F. Research Contributions

This study provides three principal contributions to clinical NLP and healthcare informatics. First, we provide the first systematic benchmark of five locally-deployed open-source LLMs (Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, Typhoon2-3.1) on Thai bilingual sports medicine EMRs, validated with Bootstrap 95% CI. This benchmark provides empirically grounded model selection guidance for privacy-preserving clinical NLP in low-resource language settings. Second, we contribute a reproducible extraction pipeline that transforms 523 unstructured Thai-English patient records into structured, machine-readable JSON format without cloud API dependency — enabling injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis. Third, we generate three key empirical insights: 1) language specialization does not confer schema-constrained extraction capability — instruction-following fidelity is the dominant determinant; 2) bilingual code-switching density is a key stratification variable for multilingual clinical NLP; and 3) inter-annotator reliability is fundamentally limited by non-standardized institutional shorthand, necessitating ICD-11 or SNOMED CT alignment.

This study demonstrated that locally-deployed 7B-parameter open-source LLMs — particularly Mistral-7B and Qwen2.5-7B — achieve robust structured clinical data extraction from Thai bilingual sports medicine EMRs (F1 > 91%) without cloud infrastructure or patient data exposure. The extraction pipeline produces machine-readable JSON output from 523 patient visit records, enabling evidence-based clinical decision support, including injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis. Annotation reproducibility remains limited by non-standardized institutional shorthand and Thai-English language divergence; adoption of ICD-11 or SNOMED CT terminology is recommended for future protocols. Language specialization does not confer a structured extraction advantage — instruction-following fidelity is the dominant performance determinant in bilingual clinical NLP settings.

VI. Conclusions

This study demonstrates that locally-deployed open-source 7B-parameter LLMs achieve clinically useful structured extraction accuracy (F1 > 91%) on bilingual Thai sports medicine EMRs without reliance on cloud infrastructure or third-party API services. Mistral-7B and Qwen2.5-7B produced statistically equivalent performance ($p \geq 0.10$), establishing both as viable candidates for privacy-preserving clinical NLP deployment. The finding that Thai-language model specialization does not confer advantage in schema-constrained bilingual extraction — and that instruction-following fidelity is the dominant performance determinant — provides actionable guidance for model selection in multilingual healthcare informatics. Annotation standardization through controlled clinical vocabulary aligned with ICD-11/SNOMED CT remains a prerequisite for improving inter-annotator reproducibility and scaling ground truth creation. The structured JSON output produced by this pipeline establishes a queryable clinical data layer enabling evidence-based decision support, including injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis — demonstrating the translational value of local LLM deployment in resource-constrained institutional healthcare settings.

Limitations of this study include: 1) inter-annotator agreement remains constrained by non-standardized institutional shorthand and Thai-English language divergence, yielding near-zero Cohen's Kappa for the Diagnosis field ($\kappa = -0.04$) and Chief Complaint field ($\kappa = -5.25$), indicating that ground truth reliability requires future validation with trained bilingual clinical annotators and standardized vocabulary; and 2) the evaluation dataset was limited to 444 of 971 available records (45.6% utilization) due to unresolvable patient ID matching, introducing potential selection bias that may affect the generalizability of the reported performance metrics. Future work should prioritize fine-tuning selected models on ICD-11- and SNOMED CT-aligned annotated corpora to improve both extraction accuracy and annotation consistency in multilingual Thai clinical settings.

Acknowledgment

This research was supported by Thammasat University Fundamental Fund, fiscal year 2026, as allocated by Thailand Science Research and Innovation (TSRI).

References

[1] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023.

[2] H. Nori et al., "Capabilities of GPT-4 on medical challenge problems," arXiv preprint arXiv:2303.13375, 2023.

[3] Office of the PDPC, "Personal Data Protection Act B.E. 2562 (2019)," Thailand, 2019.

[4] M. Agrawal et al., "Large language models are few-shot clinical information extractors," in Proc. EMNLP, 2022.

[5] D. Kartchner, S. Ramalingam, I. Al-Hussaini, O. Kronick, and C. Mitchell, "Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models," in Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, Canada, Jul. 2023, pp. 396–405. doi: 10.18653/v1/2023.bionlp-1.37.

[6] Ollama, "Run large language models locally," https://ollama.ai, 2024.

[7] J. Chen et al., "Feasibility of local LLM deployment for hospital information systems," J. Am. Med. Inform. Assoc., vol. 30, no. 9, pp. 1520-1528, 2022.

[8] V. Ntinopoulos et al., "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation," BMJ Health Care Inform., vol. 32, no. 1, e101139, 2025, doi: 10.1136/bmjhci-2024-101139.

[9] J. Huang et al., "A critical assessment of using ChatGPT for extracting structured data from clinical notes," npj Digit. Med., vol. 7, p. 106, 2024, doi: 10.1038/s41746-024-01079-8.

[10] I. C. Wiest et al., "Privacy-preserving large language models for structured medical information retrieval," npj Digit. Med., vol. 7, p. 257, 2024, doi: 10.1038/s41746-024-01233-2.

[11] L. Builtjes, J. Bosma, M. Prokop, B. van Ginneken, and A. Hering, "Leveraging open-source large language models for clinical information extraction in resource-constrained settings," JAMIA Open, vol. 8, no. 5, p. ooaf109, 2025, doi: 10.1093/jamiaopen/ooaf109.

[12] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque, "Named entity recognition in electronic health records: a methodological review," Healthc. Inform. Res., vol. 29, no. 4, pp. 286-300, 2023, doi: 10.4258/hir.2023.29.4.286.

[13] A. Kaminska et al., "Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review," Comput. Biol. Med., vol. 155, p. 106649, 2023, doi: 10.1016/j.compbiomed.2023.106649.

[14] A. Vaid, I. Landi, G. Nadkarni, and I. Nabeel, "Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders," Lancet Digit. Health, vol. 5, no. 12, pp. e855-e864, 2023, doi: 10.1016/S2589-7500(23)00202-9.

[15] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 30, e2305016120, 2023, doi: 10.1073/pnas.2305016120.

[16] K.-H. Jung, "Large language models in medicine: clinical applications, technical challenges, and ethical considerations," Healthc. Inform. Res., vol. 31, no. 2, pp. 114-124, 2025, doi: 10.4258/hir.2025.31.2.114.

[17] K. H. Lee et al., "ANNO: a general annotation tool for bilingual clinical note information extraction," Healthc. Inform. Res., vol. 28, no. 1, pp. 89-94, 2022, doi: 10.4258/hir.2022.28.1.89.