Lightweight Domain-Specific Language Model for Real-Time Structuring of Medical Prescriptions

Jonathan Pattin Cottet[1,2], Véronique Eglin[1], Alexandre Aussem[1]
[1]Universite Claude Bernard Lyon 1, CNRS, INSA Lyon,
Ecole Centrale de Lyon, LIRIS, UMR5205, 69622 Villeurbanne, France
veronique.eglin@insa-lyon.fr, alexandre.aussem@univ-lyon1.fr
[2]Phealing, Lyon, France
jonathan.pattin-cottet@phealing.fr

Abstract

Automated structuring of medical prescriptions is critical for downstream safety checks in pharmacies, yet remains challenging due to heterogeneous layouts, OCR noise, and dense clinical abbreviations in real-world documents. Existing language models either ignore layout information, rely on computationally expensive image-based architectures, or cannot operate under strict privacy and hardware constraints such as GDPR and HDS-certified environments.

We present a lightweight (<10M parameters), privacy-preserving transformer specifically designed for Entity Extraction (EE) and Entity Linking (EL) in French medical prescriptions. The model uses only OCR text and normalized 2D word coordinates, enabling robust pseudonymisation and real-time CPU-level inference while preserving essential spatial cues. It is pretrained on a large corpus of pseudonymised OCR outputs using objectives tailored to prescription structure, including a novel Token-to-Line Alignment (TLA) task, and fine-tuned on the Rx-PAD dataset (Pattin Cottet et al., 2025).

Empirical results show that our approach matches or surpasses larger document-understanding models and rivals multimodal LLMs on strict extraction metrics, while achieving sub-second latency suitable for operational deployment. The system is currently used in 230 pharmacies, demonstrating both scalability and practical relevance. These findings highlight the importance of specialized, domain-aware, lightweight models for safe, efficient, and legally compliant prescription verification.

1 Introduction

Medical prescriptions are semi-structured documents encoding dense clinical information through heterogeneous layouts, domain-specific abbreviations, and variable formatting. Extracting structured medication instructions from such documents is a prerequisite for automated downstream verification tasks, including dosage consistency checks and drug–drug interaction detection. However, this extraction remains challenging due to OCR noise, complex spatial organization, and the need to reconstruct multi-token entities into complete medication lines.

Existing NLP approaches are insufficient for this setting, because they either ignore layout or are too heavy for real-time deployment. Domain-agnostic encoder models, such as BERT (Devlin et al., 2019) or RoBERTa, process prescriptions as linear text and thus lose essential spatial information. Document-understanding models like LayoutLMv2 (Xu et al., 2021) and Donut (Kim et al., 2022) incorporate layout or image features but are computationally intensive and rely on raw images, complicating privacy-preserving deployment. Multimodal LLMs, e.g., Claude Sonnet 3.5 (Anthropic, 2024) or Pixtral (Mistral AI, 2024), provide strong few-shot reasoning but exhibit latency and cost profiles incompatible with real-time use, and are generally unsuitable for regulated healthcare environments.

To address these limitations, we propose a lightweight (<10M parameters), domain-specific transformer for Entity Extraction (EE) and Entity Linking (EL) in French medical prescriptions. The model operates solely on OCR text and normalized 2D word coordinates, enabling robust pseudonymisation, privacy compliance, and CPU-level real-time inference while preserving essential layout cues. We pretrain the model on large corpora of pseudonymised OCR outputs to capture prescription-specific patterns, and fine-tune it on the publicly available Rx-PAD dataset (Pattin Cottet et al., 2025) for structured extraction.

Our contributions are threefold:

- A privacy-preserving architecture that in-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 915–926
March 25-27, 2026

tegrates text and explicit layout information without relying on images, enabling deployment on HDS(Health Data Hosting)-compliant infrastructure.

- A domain-aware pretraining scheme that combines linguistic and spatial objectives, including a novel Token-to-Line Alignment (TLA) task, to learn prescription-specific regularities.

- A comprehensive evaluation on a publicly available labeled prescription dataset, demonstrating that a compact transformer can match or surpass larger document models and rival multimodal LLMs on strict extraction metrics, while achieving real-time CPU-level latency suitable for operational pharmacy use.

2 Use Case: Real-Time Prescription Verification in Pharmacies

In community pharmacies, structured prescription data is a prerequisite for downstream verification workflows. Pharmacists routinely consult certified drug databases to check for dosage inconsistencies, drug–drug interactions, contraindications, and mismatches between prescribed and dispensed treatments. These checks require access to structured medication instructions; however, prescriptions typically arrive as unstructured scanned documents combining typed text, handwriting, abbreviations, and provider-dependent formatting. Any automation solution must meet strict real-world constraints. Latency is critical: pharmacists may trigger extraction at any point during patient handling, and delays of more than a few seconds often lead practitioners to bypass the system. Hardware limitations also apply: most pharmacies rely on CPU-only HDS-certified servers, which preclude GPU-based models and external cloud APIs due to regulatory requirements (GDPR, data sovereignty). Image-level anonymization is insufficient, making OCR-based processing with upstream pseudonymisation of sensitive fields the only legally robust approach. Our workflow leverages the fact that prescriptions are already scanned for archiving. The scanned document is OCR-processed, pseudonymised, and passed to the domain-specific model, which extracts and links drug entities into complete medication instructions. These structured outputs feed downstream rule-based systems that perform clinical safety checks. Results are returned in real time and integrated directly into pharmacists' software, ensuring decision support without disrupting established routines. The system is currently deployed in 230 pharmacies, providing strong evidence of operational feasibility. Feedback from practitioners highlights that predictable sub-second latency and reliable structuring quality are the primary determinants of adoption, underscoring the importance of specialized, efficient models over general-purpose multimodal LLMs for this use case.

3 Related works

Knowledge Information Extraction (KIE) aims to convert unstructured documents into structured, machine-readable representations. Early transformer-based models, such as BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021), brought significant improvements in understanding sequential text through Masked Language Modeling. However, these models often struggle with semi-structured documents like medical prescriptions, where the layout and spatial relationships between text elements are critical for correct interpretation. To address this, document-aware transformer models were developed. LayoutLM (Xu et al., 2020) integrates textual content with 2D coordinates, enabling spatial reasoning, and LayoutLMv2 (Xu et al., 2021) further improves robustness by incorporating token-to-token distances and image features. GeoLayoutLM (Luo et al., 2023) introduces geometric-aware mechanisms for enhanced spatial modeling, while StructuralLM (Li et al., 2021) treats text blocks as ordered sequences to capture layout hierarchies. BROS (Hong et al., 2021) achieves layout-aware understanding using only textual content and relative positions, avoiding the computational overhead of images. For our application, real-time inference on CPU-based HDS-compliant servers and strict privacy constraints make a BROS-inspired text-and-coordinate approach particularly appealing, though we extend it with domain-specific optimizations for prescription layouts. Traditional token-labeling methods, such as the BIO scheme (Hwang et al., 2019), are effective for sequential text but face limitations when applied to semi-structured documents where entities may overlap or appear in non-linear layouts. SPADE (Hwang et al., 2021) addresses this by linking tokens within entities using a key/value chain mechanism, enhancing intra-entity connectivity. BROS (Hong et al., 2021) extends this concept to entity-relation

extraction by connecting tokens across related entities. More recently, large language models such as DocLLM (Wang et al., 2024) and LayoutLLM (Fujitake, 2024) have demonstrated impressive zero-shot and few-shot document parsing capabilities, with LayoutLLM combining spatial cues with generative reasoning to improve complex document understanding. Our approach differs by representing all tokens as nodes in a fully connected undirected graph, allowing simultaneous Entity Extraction (EE) and Entity Linking (EL). Unlike SPADE (Hwang et al., 2021), which focuses primarily on local spatial chains, and BROS (Hong et al., 2021) , which is limited to single relation types, our model can handle the intricate, overlapping structures of medical prescriptions. This enables robust parsing of 61 entity types and 11 relation types, making it well-suited for real-world pharmacy applications where accurate and real-time information extraction is essential.

4 Methodology

4.1 Overview

We propose a lightweight, domain-specific language model for extracting and linking entities from French medical prescriptions in real-time. The model is trained from scratch using OCR-extracted text and 2D word-level positions. Unlike pre-trained models such as CamemBERT-bio (Touchent et al., 2023), our approach avoids sequence length constraints and irrelevant vocabulary, enabling efficient learning on domain-specific layouts and terminology.

To enhance layout comprehension and robustness to noisy OCR outputs, we introduce a Token-to-Line Alignment (TLA) pretraining objective. In this task, each token is supervised to predict the line in the prescription to which it belongs, as detected by OCR. The model receives both the token embeddings and normalized 2D coordinates and is trained to assign tokens to the correct line group, even when scans are degraded or handwriting is unclear. TLA complements masked language modeling (MLM) and area-masked LM (AMLM) objectives by explicitly encoding the spatial structure of prescriptions. This encourages the model to capture token dependencies within the same medication instruction, improving the reconstruction of multi-token entities into complete, structured medication lines during downstream Entity Extraction (EE) and Entity Linking (EL) tasks.

4.2 Language Model Pre-training

4.2.1 Data

Pre-training uses 330,000 anonymized prescriptions collected from partner pharmacies. OCR text is pseudonymized according to CNIL recommendations (CNIL, 2019), replacing identifiers irreversibly on the host server. Only pseudonymized OCR outputs are used for model training. We measured OCR performance on 100 prescriptions: character error rate <1%, and line-creation errors 5%. These minor errors justify our word-level approach and TLA task.

4.2.2 Tokenizer and Preprocessing

We train a byte-level BPE tokenizer on a subset of 100k pseudonymised OCR prescriptions, with a vocabulary of 5,002 tokens specifically designed to cover common drug names, pathologies, dosages, and medical devices. Custom pre-tokenization rules preserve meaningful entity boundaries; for example, “500mg/10mg” is split into two distinct tokens to maintain the integrity of dosage information. Out-of-vocabulary issues are mitigated by the byte-level encoding, which guarantees that any string can be decomposed into valid subword units. Token sequences are capped at 600 tokens during pretraining to balance full coverage of prescription content with efficient memory and computation requirements. Prescriptions exceeding this limit are truncated by retaining the first 600 tokens, which correspond to the highest-density information regions in practice. This affects less than 3% of samples in our corpus and does not degrade downstream DAC performance. At inference time, prescriptions are processed individually and are not subject to this sequence-length constraint.

4.2.3 Spatial Language Model Architecture

Architecturally, our model is a from-scratch implementation of a layout-aware transformer encoder. While it adopts a spatial-aware attention mechanism inspired by BROS (Hong et al., 2021), all 7.6M parameters are initialized randomly. This approach allows us to fully customize the 8-layer, 256-hidden-unit architecture for the specific linguistic and spatial regularities of medical prescriptions without being constrained by the pre-existing weights or vocabularies of domain-agnostic models. Each token is represented by its embedding and four vertex coordinates $(P_{tl}, P_{tr}, P_{br}, P_{bl})$, normalized in $[0, 1]$. Relative positional encoding is

computed from pairwise token distances and integrated into the attention mechanism:

$$a_{i,j}^{h} = (W_h^q t_i)^T (W_h^k t_j) + (W_h^q t_i)^T \vec{b}_{i,j} \quad (1)$$

Pre-training uses three objectives jointly: MLM (Devlin et al., 2019), AMLM (Hong et al., 2021), and TLA. The TLA task encourages learning token-to-line associations using OCR-detected lines as supervision.

4.3 LM Fine-tuning for Task-specific Objectives

All LM layers are unfrozen during fine-tuning, with three task-specific heads: one for entity extraction (EE) and two for entity linking (EL) (see Fig. 1). EE predicts one of 61 entity types per token, including drug names, dosages, routes, patient info, and prescriber details. EL first predicts token group membership via multi-label classification, then links tokens using a dot-product adjacency matrix with a 0.5 threshold. Heads are trained jointly.

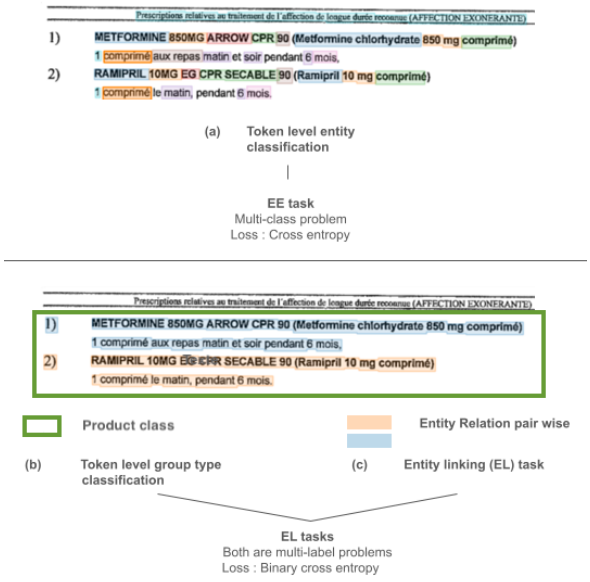


Figure 1: Top: EE head; Bottom: two EL heads.

5 Experiments

5.1 Dataset

We evaluate our model on a public anonymized dataset of 200 French medical prescriptions for fine-tuning and testing, covering 61 EE tags and 11 EL groups. The dataset is split into 150 training and 50 evaluation samples. Annotations were performed by three annotators with quality control and pharmacist review. This dataset is publicly available as Rx-PAD (Pattin Cottet et al., 2025). This setup enables reproducible benchmarking while maintaining anonymity."

5.2 Implementation

5.2.1 Pre-training

The model uses an 8-layer transformer with 256 hidden units and 4 attention heads. OCR is performed via a Health Data Hosting (HDS) certified service. Pre-training runs for 45 epochs with a batch size of 150, optimized with Adam and learning rate scheduling.

5.2.2 Fine-tuning

Fine-tuning is performed on EE and EL tasks for 400 epochs, batch size 60, with learning rate decay from $1E{-}4$ to $1E{-}5$. Post-processing uses a graph-based approach to infer missing links. Evaluation metrics include EE-F1, EL-F1, and DAC (Drug Accuracy and Completeness), which measures the proportion of prescriptions where all three key elements—drug name, dosage, and form—are correctly extracted and linked, reflecting end-to-end extraction quality.

5.3 Results

Model	EE-F1	EL-F1	DAC
LM (ours, random init)	85.62	83.41	88.3
LM (C-Bio init)	82.36	78.8	–
LayoutLM2 Finetuned	68.17	–	–
BROS Finetuned	67.29	–	–
NOVA PRO Zero-shot	–	–	74.3
CS 3.5 Zero-shot	–	–	88.4
NOVA PRO 3-shot	–	–	78.6
CS 3.5 3-shot	–	–	88.7

Table 1: Performance on the public prescription dataset used for evaluation. DAC measures complete correctness of drug name, form, and dosage for each prescription.

Model	Inference Time	Cost	#Params
LM (ours)	0.79s	$0.002	8M
NOVA PRO	36s	$0.006	90B
CS 3.5	42s	$0.01	400B

Table 2: Inference latency and cost per document on the public prescription evaluation dataset.

EL-F1 is not reported for LayoutLMv2 and BROS because these models do not natively support our multi-relation, overlapping entity-linking

Pre-training Objectives	EE-F1	EL-F1	DAC
MLM + AMLM (Baseline)	70.72	65.33	71.6
MLM + AMLM + Zone Prediction	71.91	67.79	73.2
MLM + AMLM + TLA (Ours)	**76.39**	**72.89**	**77.1**

Table 3: Ablation study on pre-training objectives. TLA yields the largest gains in EL-F1 and DAC by explicitly modeling prescription line structure. Results are reported after 1 epoch of pre-training and 80 epochs of fine-tuning on Rx-PAD. EE-F1/EL-F1: F1-scores; DAC: Drug Accuracy and Completeness.

formulation without substantial architectural modification. Moreover, EL-F1 is conditional on correct entity extraction: when EE-F1 is substantially lower, meaningful entity linking is not achievable. For this reason, we do not report EL results for models whose EE performance is substantially lower, as meaningful entity linking presupposes reliable entity extraction. Similarly, zero-shot and few-shot LLM baselines produce free-form text rather than explicit entity graphs and are therefore evaluated only via the end-to-end DAC metric.

Common errors primarily arise from ambiguous abbreviations, misclassification of dosage forms, and grouping failures. For example, abbreviations can be misleading: laboratory names or units may resemble dosage units, causing incorrect token labeling. Another frequent issue is the improper grouping of tokens into medication lines, which can result in incomplete or fragmented entity linking. While such errors may slightly impact EE- and EL-F1 scores, the DAC metric demonstrates that the majority of medication lines are correctly extracted and fully structured. This emphasizes that, despite minor token-level errors, the model reliably produces end-to-end medication information suitable for real-time pharmacy workflows.

DAC is particularly important in practical pharmacy settings because it evaluates the correctness of an entire medication instruction as a unit, rather than individual tokens or entity links. Unlike EE- or EL-F1 scores, which may still be high even if a prescription is partially incorrect, DAC captures whether pharmacists can safely rely on the structured output without manually verifying each field. High DAC scores indicate that the model produces fully usable, end-to-end prescription representations, aligning directly with the operational goal of reducing verification time and minimizing human error. While a DAC below 90% would be insufficient for unsupervised use, in practice the system is designed as a decision-support tool: pharmacists always retain final control. Error rates above 10% would be clinically unacceptable for full automation, but are acceptable in assistive settings where outputs are reviewed.

Claude Sonnet 3.5 benefits from access to the full prescription image, which allows it to leverage visual cues such as alignment, spacing, and handwriting structure. This is particularly helpful for noisy or tilted scans, where text lines may not be perfectly segmented by OCR. As a result, Claude more reliably groups tokens belonging to the same medication instruction and avoids mixing information across drugs. This leads to slightly higher DAC in few-shot settings due to its strong generative reasoning and implicit world knowledge, which can correct minor OCR and formatting inconsistencies. However, this comes at the cost of high latency, operational expense, and limited deployment feasibility in regulated pharmacy environments.

Overall, our model achieves high accuracy and low latency, making it suitable for real-time deployment in pharmacies, unlike large multimodal LLMs (Anthropic, 2024; Mistral AI, 2024) which are slower and costlier.

5.4 Ablation Study

To evaluate the contribution of our proposed Token-to-Line Alignment (TLA) task, we conducted an ablation study comparing it against standard pre-training objectives. This test isolates the impact of spatial supervision on the model's ability to reconstruct prescription structures. For efficiency, all models in this study were pre-trained for only 1 epoch on the private dataset and fine-tuned for 80 epochs on the Rx-PAD dataset (Pattin Cottet et al., 2025).

As shown in Table 3, while adding a generic "Zone Prediction" task (classifying tokens into document regions) offers incremental improvements, the TLA objective provides a significant performance boost across all metrics. Notably, TLA increases the DAC score by +5.5 points over the baseline (MLM + AMLM), demonstrating that explicitly modeling line-level associations is superior to general spatial tasks for capturing the unique regularities of medical prescriptions.

6 Conclusion & Future Work

We presented a lightweight, domain-specific language model for real-time entity extraction and linking in French medical prescriptions. Deployment in 230 pharmacies demonstrates its practical value: the model reliably assists pharmacists in verifying prescriptions, detecting errors, and ensuring safe dispensation. Our results show that task-specific architectures can meet critical constraints of accuracy, latency, cost-efficiency, and regulatory compliance, where general-purpose models often fall short.

Although multimodal LLMs such as Claude Sonnet 3.5 (Anthropic, 2024) achieve high accuracy, their latency and operational cost limit real-time pharmacy use. Future work may explore hybrid approaches that combine the reasoning capabilities of LLMs with the efficiency of domain-specific layout-aware models, inspired by designs like LayoutLLM (Fujitake, 2024). Such solutions could merge the semantic flexibility of large models with the speed, privacy, and cost-effectiveness of specialized encoders.

These results underscore the value of domain-specific pre-training, layout-aware modeling, and lightweight architectures for mission-critical healthcare applications, providing a path for integrating advanced AI reasoning into practical, large-scale deployment.

Limitations

Several limitations of our work should be noted. First, large language models (LLMs) exhibit slower inference times, which hinders real-time deployment in pharmacy settings. Future research could explore optimization strategies that balance speed and accuracy, such as fine-tuning smaller LLMs on HDS-compliant servers, output compression combined with algorithmic reconstruction, or hybrid multimodal approaches that enhance performance without compromising accuracy or user trust.

Second, our reliance on OCR text means that extreme degradation in scan quality or highly unconventional handwriting can impact the upstream tokenization and 2D coordinate accuracy. While the TLA task mitigates minor line-creation errors, the system is most robust on documents with legible text headers and typed instructions.

Third, our current evaluation lacks extensive feedback from practicing pharmacists. Incorporating practitioner input would provide valuable insights into recurrent drug extraction errors, usability issues, and workflow integration, helping refine the system based on real-world usage patterns.

Finally, our model is tailored specifically to French prescriptions. Differences in prescription formats, terminology, and layout conventions across languages and healthcare systems limit direct applicability elsewhere. Future work should investigate cross-lingual adaptability and design methods that accommodate region-specific prescription practices, broadening the model's utility in international settings.

Ethical Considerations

Processing medical prescriptions entails handling highly sensitive personal health information. Our approach complies with GDPR and relevant healthcare data protection standards by training the language model exclusively on OCR-extracted text that has been pseudonymised using a CNIL-approved procedure. Sensitive fields—including names, phone numbers, and identification numbers—are irreversibly anonymised to eliminate any risk of re-identification, while preserving the linguistic and structural characteristics necessary for model training.

Pseudonymisation at the image level is considerably more complex and less reliable, further motivating our text-based approach. This strategy ensures both strong legal compliance and practical efficiency in handling sensitive medical data.

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-03-15.

CNIL. 2019. Pseudonymisation et anonymisation des données. Accessed: 2025-11-18.

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

Masato Fujitake. 2024. LayoutLLM: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10219–10224, Torino, Italia. ELRA and ICCL.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. BROS: A Pre-trained Language Model for Understanding Texts in Document.

Wonjun Hwang, Sungrae Park, Byeonggeun Lee, and et al. 2019. Post-OCR Parsing: Building Simple and Robust Parser via BIO Tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.

Wonjun Hwang, Jeonghyeon Yim, Sangho Park, and et al. 2021. Spatial Dependency Parsing for Semi-Structured Document Information Extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343.

Geewook Kim, Teakgyu Hong, Moonbin Kim, Junyeop Kim, and Gunhee Han. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*. Springer.

Chang Li, Bin Bi, Ming Yan, Weizhu Wang, Shuming Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural Pre-training for Form Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6309–6318.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. *Preprint*, arXiv:2304.10759.

Mistral AI. 2024. Pixtral: Vision-language model by mistral ai. `https://mistral.ai/news/pixtral-release`. Accessed: 2025-03-15.

Jonathan Pattin Cottet, Vincent Eglin, and Alex Aussem. 2025. Rx-pad: Recognition and extraction – a dataset for prescription analysis and clinical data structuring. In *Document Analysis and Recognition – ICDAR 2025*, volume 16027 of *Lecture Notes in Computer Science*, pages 145–160, Cham. Springer.

Rian Touchent, Laurent Romary, and Eric De La Clergerie. 2023. CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 323–334, Paris, France. ATALA.

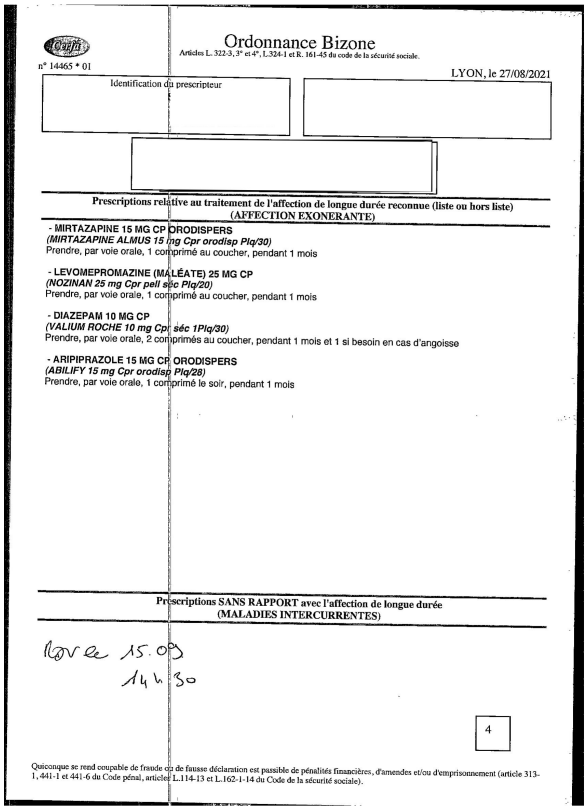
Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1192–1200.

Yiheng Xu, Yang Xu, Tengchao Lv, Lei Cui, Furu Wei, Guangyan Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Weizhu Chen et al. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2579–2591.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Additional Figures from the Public Prescription Dataset



Ordonnance Bizone

Articles L. 322-3, 3° et 4°, L.324-1 et R. 161-45 du code de la sécurité sociale.

n° 14465 * 01

LYON, le 27/08/2021

Identification du prescripteur

Prescriptions relative au traitement de l'affection de longue durée reconnue (liste ou hors liste) (AFFECTION EXONERANTE)

- MIRTAZAPINE 15 MG CP ORODISPERS
(MIRTAZAPINE ALMUS 15 mg Cpr orodisp Plq/30)
Prendre, par voie orale, 1 comprimé au coucher, pendant 1 mois

- LEVOMEPROMAZINE (MALÉATE) 25 MG CP
(NOZINAN 25 mg Cpr pell séc Plq/20)
Prendre, par voie orale, 1 comprimé au coucher, pendant 1 mois

- DIAZEPAM 10 MG CP
(VALIUM ROCHE 10 mg Cpr séc 1Plq/30)
Prendre, par voie orale, 2 comprimés au coucher, pendant 1 mois et 1 si besoin en cas d'angoisse

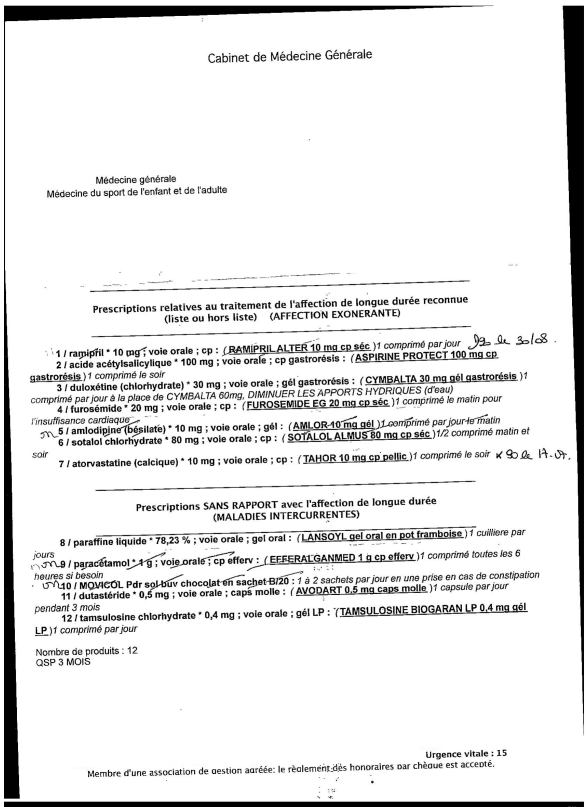
- ARIPIPRAZOLE 15 MG CP ORODISPERS
(ABILIFY 15 mg Cpr orodisp Plq/28)
Prendre, par voie orale, 1 comprimé le soir, pendant 1 mois

Prescriptions SANS RAPPORT avec l'affection de longue durée (MALADIES INTERCURRENTES)

Revue 15.09
14 h 30

4

Quiconque se rend coupable de fraude ou de fausse déclaration est passible de pénalités financières, d'amendes et/ou d'emprisonnement (article 313-1, 441-1 et 441-6 du Code pénal, articles L.114-13 et L.162-1-14 du Code de la sécurité sociale).



Cabinet de Médecine Générale

Médecine générale
Médecine du sport de l'enfant et de l'adulte

Prescriptions relatives au traitement de l'affection de longue durée reconnue (liste ou hors liste) (AFFECTION EXONERANTE)

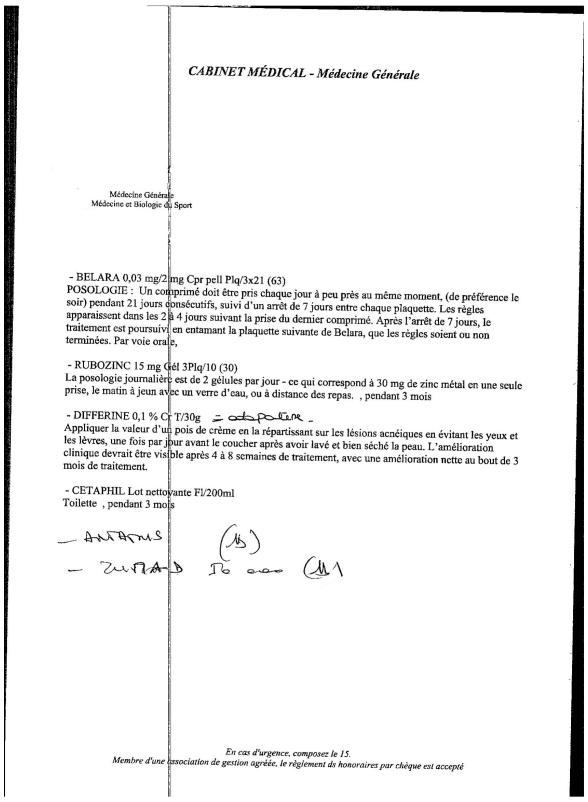
1 / rapipril * 10 mg ; voie orale ; cp : (RAMIPRIL ALTER 10 mg cp séc) *1 comprimé par jour* Rdv le 30/08
2 / acide acétylsalicylique * 100 mg ; voie orale ; cp gastrorésis : (ASPIRINE PROTECT 100 mg cp gastrorésis) *1 comprimé le soir*
3 / duloxétine (chlorhydrate) * 30 mg ; voie orale ; gél gastrorésis : (CYMBALTA 30 mg gél gastrorésis) *1 comprimé par jour à la place de CYMBALTA 60mg, DIMINUER LES APPORTS HYDRIQUES (d'eau)*
4 / furosémide * 20 mg ; voie orale ; cp : (FUROSEMIDE EG 20 mg cp séc) *1 comprimé le matin pour l'insuffisance cardiaque*
5 / amlodipine (bésilate) * 10 mg ; voie orale ; gél : (AMLOR 10 mg gél) *1 comprimé par jour le matin*
6 / sotalol chlorhydrate * 80 mg ; voie orale ; cp : (SOTALOL ALMUS 80 mg cp séc) *1/2 comprimé matin et soir*
7 / atorvastatine (calcique) * 10 mg ; voie orale ; cp : (TAHOR 10 mg cp pellic) *1 comprimé le soir* K 90 du 17.09.

Prescriptions SANS RAPPORT avec l'affection de longue durée (MALADIES INTERCURRENTES)

8 / paraffine liquide * 78,23 % ; voie orale ; gel oral : (LANSOYL gel oral en pot framboise) *1 cuillière par jours*
9 / paracétamol * 1 g ; voie orale ; cp efferv : (EFFERALGANMED 1 g cp efferv) *1 comprimé toutes les 6 heures si besoin*
10 / MOVICOL Pdr sol buv chocolat en sachet B/20 : *1 à 2 sachets par jour en une prise en cas de constipation*
11 / dutastéride * 0,5 mg ; voie orale ; caps molle : (AVODART 0,5 mg caps molle) *1 capsule par jour pendant 3 mois*
12 / tamsulosine chlorhydrate * 0,4 mg ; voie orale ; gél LP : (TAMSULOSINE BIOGARAN LP 0,4 mg gél LP) *1 comprimé par jour*

Nombre de produits : 12
QSP 3 MOIS

Urgence vitale : 15
Membre d'une association de gestion agréée: le règlement des honoraires par chèque est accepté.



CABINET MÉDICAL - Médecine Générale

Médecine Générale
Médecine et Biologie du Sport

- BELARA 0,03 mg/2 mg Cpr pell Plq/3x21 (63)
POSOLOGIE : Un comprimé doit être pris chaque jour à peu près au même moment, (de préférence le soir) pendant 21 jours consécutifs, suivi d'un arrêt de 7 jours entre chaque plaquette. Les règles apparaissent dans les 2 à 4 jours suivant la prise du dernier comprimé. Après l'arrêt de 7 jours, le traitement est poursuivi en entamant la plaquette suivante de Belara, que les règles soient ou non terminées. Par voie orale,

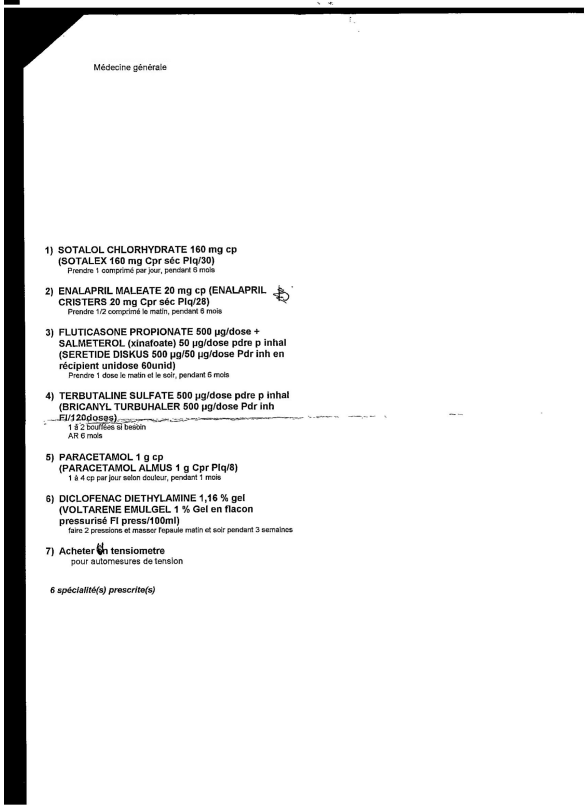
- RUBOZINC 15 mg Gél 3Plq/10 (30)
La posologie journalière est de 2 gélules par jour - ce qui correspond à 30 mg de zinc métal en une seule prise, le matin à jeun avec un verre d'eau, ou à distance des repas. , pendant 3 mois

- DIFFERINE 0,1 % Cr T/30g = adapolène
Appliquer la valeur d'un pois de crème en la répartissant sur les lésions acnéiques en évitant les yeux et les lèvres, une fois par jour avant le coucher après avoir lavé et bien séché la peau. L'amélioration clinique devrait être visible après 4 à 8 semaines de traitement, avec une amélioration nette au bout de 3 mois de traitement.

- CETAPHIL Lot nettoyante Fl/200ml
Toilette , pendant 3 mois

— ANTARUS (AB)
— ZYRTEC 10 mg (AM)

En cas d'urgence, composez le 15.
Membre d'une association de gestion agréée, le règlement de honoraires par chèque est accepté



Médecine générale

1) **SOTALOL CHLORHYDRATE 160 mg cp**
(SOTALEX 160 mg Cpr séc Plq/30)
Prendre 1 comprimé par jour, pendant 6 mois

2) **ENALAPRIL MALEATE 20 mg cp (ENALAPRIL CRISTERS 20 mg Cpr séc Plq/28)**
Prendre 1/2 comprimé le matin, pendant 6 mois

3) **FLUTICASONE PROPIONATE 500 µg/dose + SALMETEROL (xinafoate) 50 µg/dose pdre p inhal (SERETIDE DISKUS 500 µg/50 µg/dose Pdr inh en récipient unidose 60unid)**
Prendre 1 dose le matin et le soir, pendant 6 mois

4) **TERBUTALINE SULFATE 500 µg/dose pdre p inhal (BRICANYL TURBUHALER 500 µg/dose Pdr inh Fl/120doses)**
1 à 2 bouffées si besoin
AR 6 mois

5) **PARACETAMOL 1 g cp**
(PARACETAMOL ALMUS 1 g Cpr Plq/8)
1 à 4 cp par jour selon douleur, pendant 1 mois

6) **DICLOFENAC DIETHYLAMINE 1,16 % gel (VOLTARENE EMULGEL 1 % Gel en flacon pressurisé Fl press/100ml)**
faire 2 pressions et masser l'epaule matin et soir pendant 3 semaines

7) **Acheter un tensiometre**
pour automesures de tension

6 spécialité(s) prescrite(s)

Figure 2: Sample prescriptions from the publicly available fine-tuning dataset (authors anonymized).

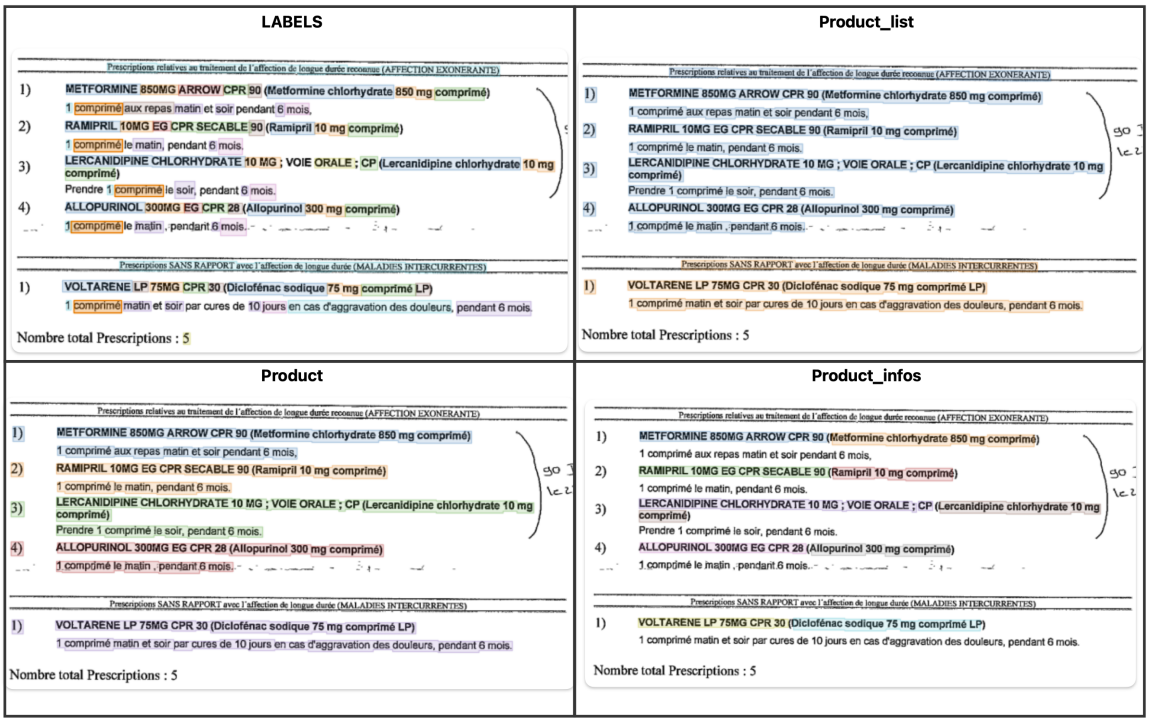


Figure 3: Example of one prescription from the publicly available fine-tuning dataset (authors anonymized) illustrating the hierarchical structure used to model complex relationships between entities. The top-left section shows all labeled entities used for Entity Extraction (EE), while the remaining sections highlight different types of entity groupings used for Entity Linking (EL). Colors indicate shared group IDs within each group type, showing how entities can participate in multiple overlapping structures.

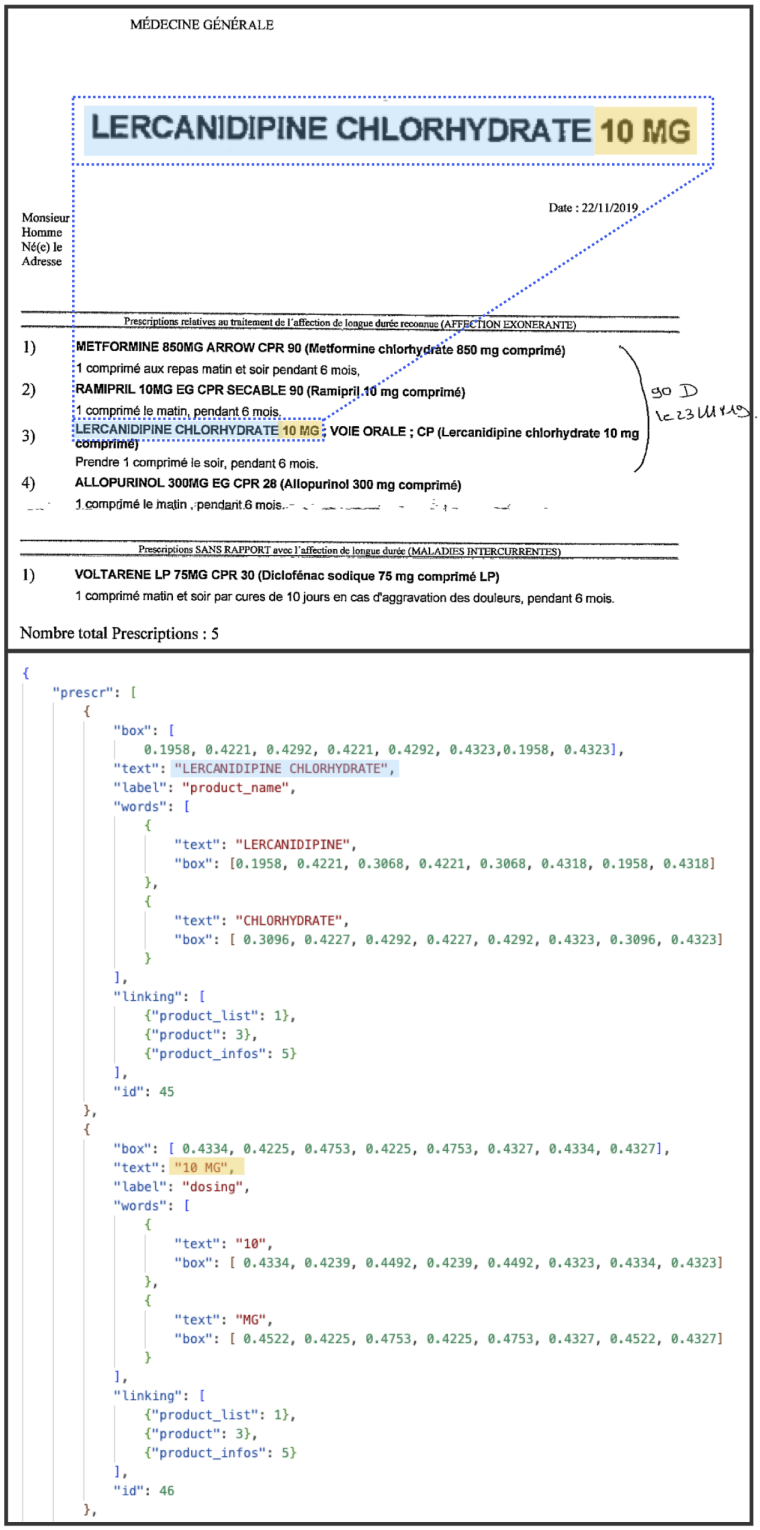


Figure 4: Example of a prescription showing its scanned image alongside its labeled JSON format, illustrating how the model represents and structures entities for downstream tasks.

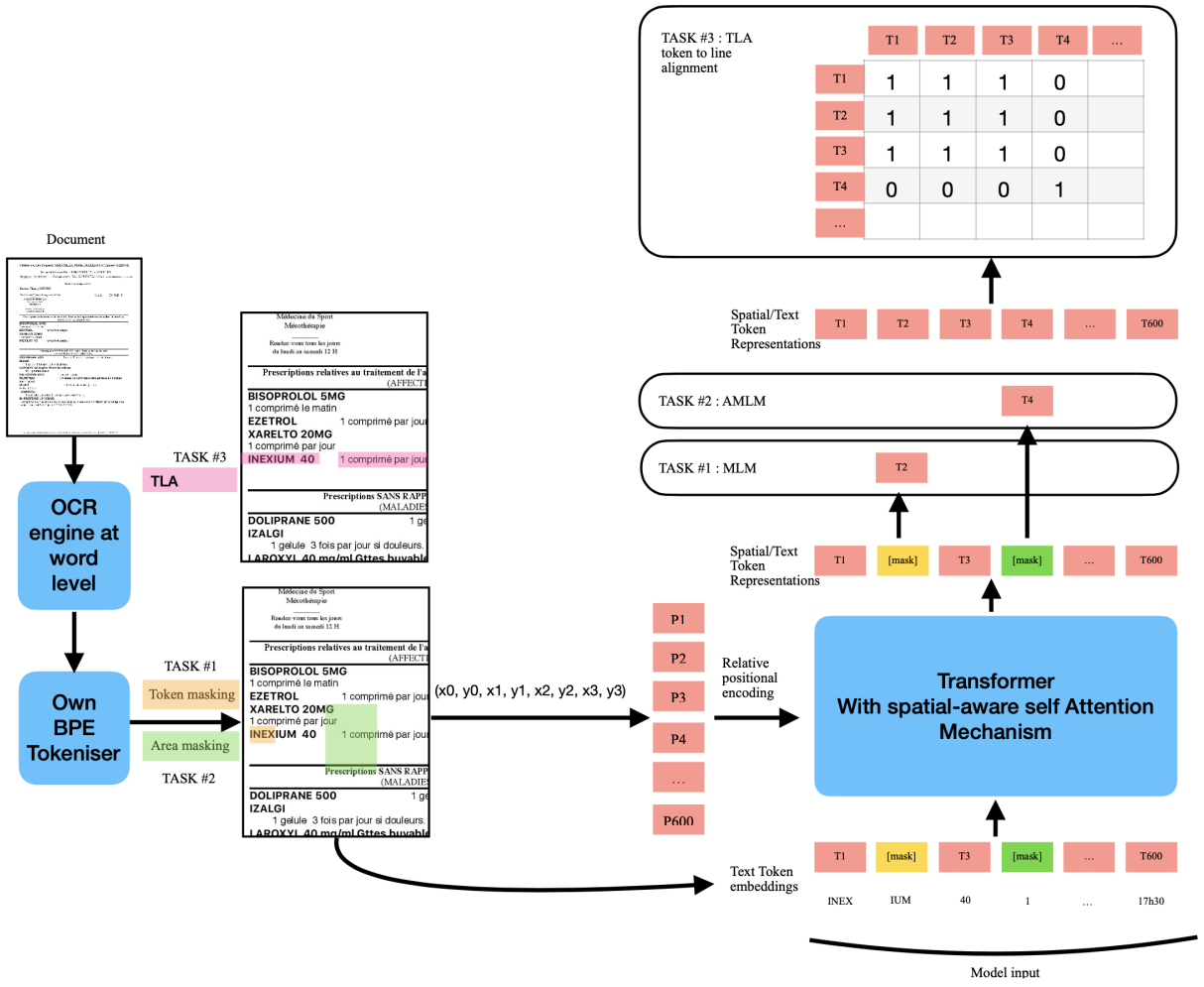


Figure 5: Visual description of our language model pre-training tasks. Task 1 corresponds to Masked Language Modeling (MLM), Task 2 to Area-Masked Language Modeling (AMLM), and Task 3 to Token-to-Line Alignment (TLA).

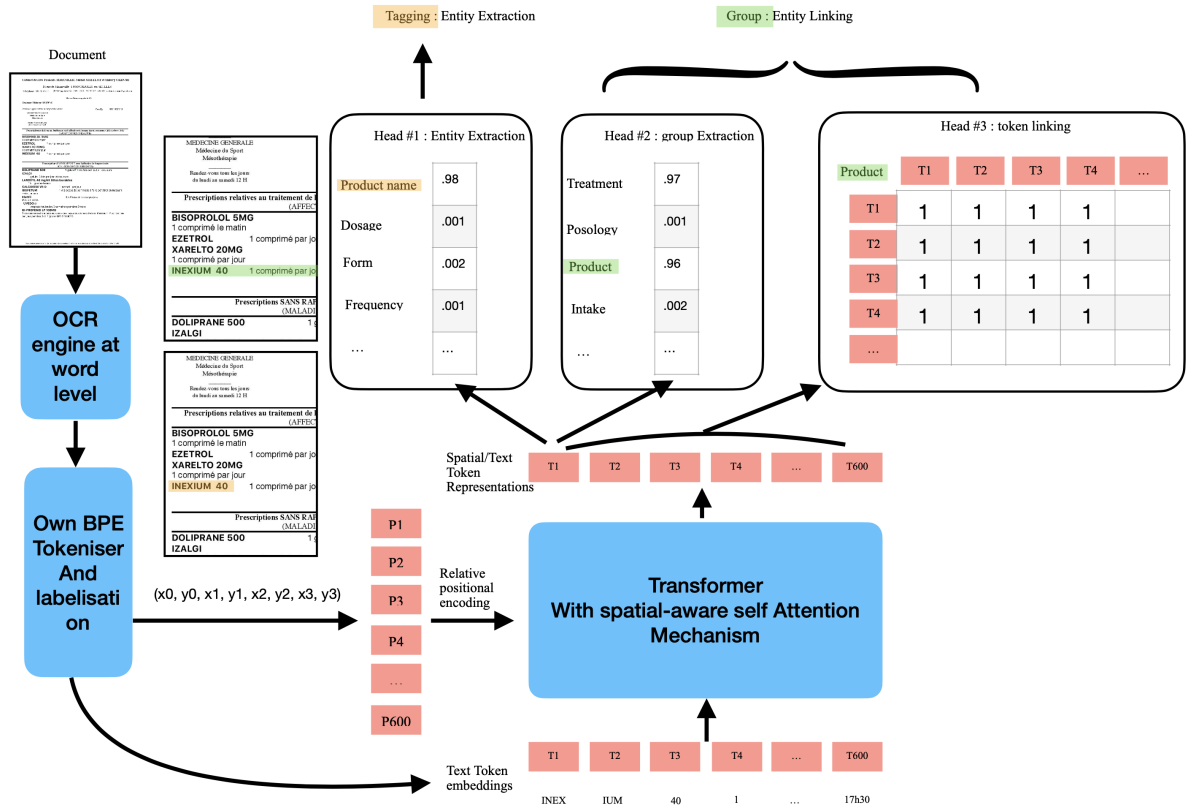


Figure 6: Visual description of our model for fine-tuning, including downstream parser tasks. For Entity Extraction (head #1), the model performs a token-level multi-class classification. For linking these entities, the parser combines the two subtasks (head #2 and head #3) to correctly link tokens belonging to the same group.