Thesis Outline

Public-Data Pretraining for Clinical Information Extraction

Rian Touchent

January 2025

Sorbonne Université / INRIA Paris (ALMAnaCH)

Outline

Introduction

Part 1: Building a Biomedical Corpus

Part 2: Pretraining Language Models

Part 3: Adapting to Clinical Tasks

Timeline

Introduction

Introduction

- Dominant paradigm: pretraining at scale
- Healthcare limitation: clinical data is confidential and scarce
- Question: can we exploit public data for the clinical domain?

Related Works

1. Language Models
2. Corpus Annotation
3. Clinical Information Extraction

Part 1: Building a Biomedical Corpus

Where to find public biomedical text?

Ch.1: Collecting Biomedical Text

Publication: CamemBERT-bio

biomed-fr: First public French biomedical corpus

Source	Content	Size
ISTEX	Scientific literature	276M words
CLEAR	Drug leaflets	73M words
E3C	Clinical cases & leaflets	64M words
Total		**413M words**

→ Only 2.7 GB (vs 138 GB OSCAR)

Ch.1: NER Results

Result: +2.5 F1 avg across 5 benchmarks

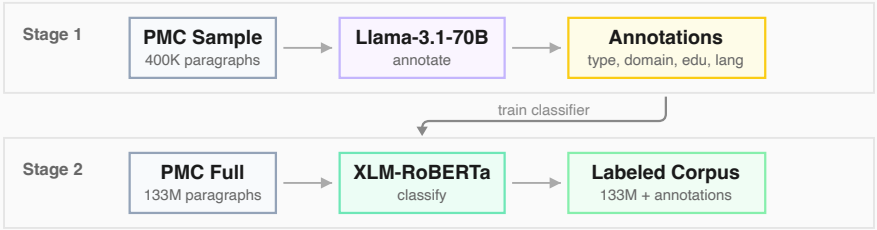
Style	Dataset	CamemBERT	CamemBERT-bio
Clinical	CAS1	70.50	**73.03**
Clinical	CAS2	79.02	**81.66**
Clinical	E3C	67.63	**69.85**
Leaflets	EMEA	74.14	**76.71**
Scientific	MEDLINE	65.73	**68.47**
Average		71.40	**73.94**

Is all this text useful? Is there clinical content hidden?

Ch.2: Detecting Content Types

Publication: Biomed-Enriched

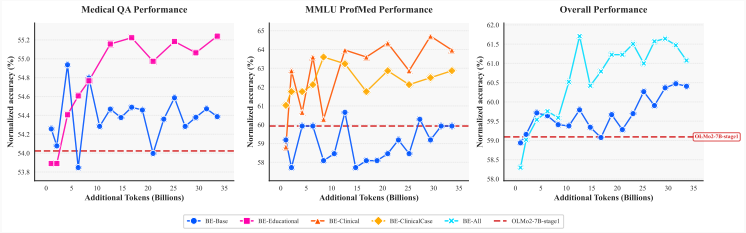
Problem: 91.6% of PMC articles mix content types



Result: Extract **2M clinical case paragraphs** from PMC

Ch.2: Biomed-Enriched Results

Finding: Same performance with 1/3 of tokens



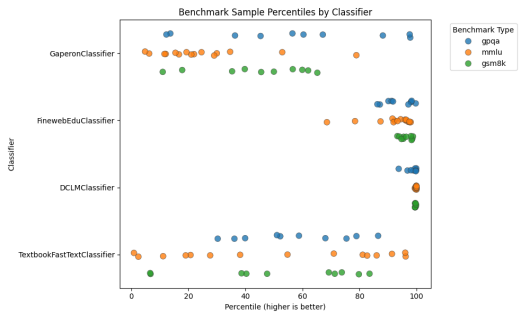
- BE-All reaches target at 12.6B tokens (vs 33.6B baseline)
- 77.3% F1 ≈ BioClinical-ModernBERT with 2.5× fewer tokens

Ch.2: But Quality Filtering is Risky

Publication: GAPeron – BiaHS contribution

BiaHS (Benchmark-in-a-Haystack):

- Inject 35 benchmark samples in 100k docs
- Test: where do classifiers rank them?



Finding: DCLM ranks MMLU/GSM8K in top-5% $\rightarrow$ 20$\times$ amplification

$\rightarrow$ GAPeron classifier (general quality) does NOT amplify

Can we build a better French biomedical corpus?

Ch.3: MC-Bio Corpus – Quality Signals

Fine-grained annotation (Qwen3-235B on 2.16M paragraphs)

Quality signals (1-10):

Signal	Mean
educational_score	6.5
content_richness	6.8
writing_quality	7.6
terminology_precision	7.6

Content types (12):

Type	Count
research_findings	526k
medical_knowledge	388k
clinical_guidance	218k
patient_case	36k
...	...

Hard filters: exclude `other`, `drug_information`, len < 50

Ch.3: Content Type Ablation

Which content types help? (remove one, measure Δ vs baseline)

Removed Content Type	Avg F1	Δ
drug_information	64.35%	**+0.67pp**
research_findings	64.24%	+0.56pp
policy_administrative	64.07%	+0.39pp
random (control)	63.72%	+0.04pp
all (baseline)	63.68%	ref
medical_knowledge	63.55%	−0.13pp
research_methodology	63.51%	−0.17pp

→ **Exclude** drug_information (+0.67pp) – redundant with EMEA

Ch.3: Quality Signal Ablation

Which signals help? (threshold ≥7)

Threshold ablation:

Threshold	Avg F1	Δ
edu≥8 AND cont≥8	59.37%	+0.55pp
edu≥7 AND cont≥7	**59.30%**	**+0.48pp**
edu≥6 AND cont≥6	58.95%	+0.13pp
baseline (no filter)	58.82%	ref

Individual signals:

Signal	Avg F1	Δ
edu≥7 AND cont≥7	59.30%	+0.48pp
edu≥7	59.04%	+0.22pp
cont≥7	59.02%	+0.20pp
baseline	58.82%	ref
term≥7	58.57%	−0.26pp
writ≥7	58.49%	−0.33pp

→ **edu + cont synergistic** (+0.48pp), term/writ hurt

Ch.3: FineWeb-Style Upsampling

Quality-weighted sampling per article

Quality Ratio	Bucket	Coefficient
0%	excluded	0×
1–25%	low	4.3×
26–50%	medium	8.5×
51–75%	high	12.8×
76–99%	very high	21.3×
100%	perfect	34.1×

Quality ratio = paragraphs with (edu ≥ 7 AND content ≥ 7) / total

Ch.3: MC-Bio Recipe (10B tokens)

Source	Tokens	%	Content
MC-Bio (quality-weighted)	7B	70%	Medical knowledge
MCQA synthetic (Ch.6)	2B	20%	QA pairs
E3C (clinical sentences)	400M	4%	Patient cases
EMEA (drug notices)	600M	6%	Pharmacology
Total	**10B**	100%	

→ Used to train ModernCamemBERT-bio (Ch.5)

Part 2: Pretraining Language Models

How to use the corpus to adapt a model?

Ch.4: Domain Adaptation

Publication: CamemBERT-bio

Approach: Continual pretraining from CamemBERT

- Start from general French model weights
- Continue MLM on biomed-fr corpus
- 50k steps, 39h on 2× V100

Result: Simple and efficient

- +2.5 F1 avg on biomedical NER
- Public model usable by all hospitals

Finding: Continual pretraining is 32× greener

Model	GPU-hours	Hardware	CO_2 (kg)
DrBERT	2,560	128× V100	26.11
AliBERT	960	48× A100	8.16
CamemBERT-bio	**78**	2× V100	**0.8**

→ Contradicts DrBERT claim that continual pretraining doesn't work

Architectures have evolved.

Can we leverage Flash Attention and long context for clinical documents?

Ch.5: ModernCamemBERT-bio – Motivation

Publication: ModernCamemBERT-bio

Problem: Clinical reports are long

- ICD-10 coding (FRACCO): full discharge summaries
- BERT context: 512 tokens $\rightarrow$ truncation

Modern architecture:

- Flash Attention $\rightarrow$ efficient long sequences
- RoPE $\rightarrow$ better position encoding
- 8,192 token context (16$\times$ BERT)

Question: For pretraining, CLM or MLM?

Ch.5: The CLM Detour

Idea: Two-phase training

Why CLM first?

- CLM: Every token predicts next → uniform gradient signal
- MLM: Only 15% masked → sparse gradient signal

Gradient CV: CLM = 0.12 (uniform) vs MLM = 0.59 (sparse)

Ch.5: CLM Improves ALL Tokens

Key ablation: Pooling strategy comparison

Pooling	CLM F1	MLM F1	Gap
CLS token	23.0%	18.1%	+4.9pp
Mean pooling	**39.9%**	32.8%	**+7.1pp**

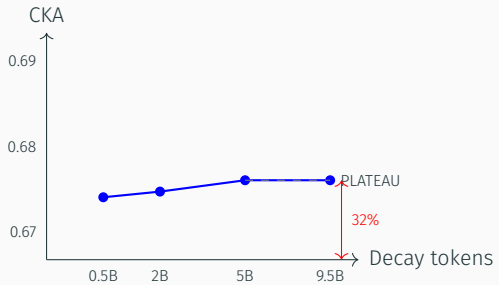
Finding: Gap is LARGER with mean pooling (+7.1pp vs +4.9pp)

→ CLS concentration is NOT the mechanism

→ CLM improves **all token representations**, not just CLS

Ch.5: Computational Hysteresis

Question: Does MLM decay undo CLM compression?



No. CKA similarity (CLM→decay vs MLM):

- 0.5B decay: 0.676
- 9.5B decay: 0.679
- Δ = +0.003 (plateau)

32% permanent divergence

→ CLM effects persist despite MLM decay

Experiment: Transplant CLM layers into MLM model

Transplant	F1	% of gap recovered
+ early (L0–7)	31.9%	−7% (worse!)
+ mid (L8–14)	35.3%	**+37%**
+ late (L15–21)	32.4%	−1% (worse!)
+ all attention	33.9%	+19%
+ all MLP	35.3%	+37%
Max recovery:		**37%**

→ No single component captures >37% of the advantage

→ The effect is **emergent**: distributed across the entire network

Ch.5: MLM Can't Exploit Rich Data

Control experiment: Same MCQA data, different objective

Model	Objective	Δ vs baseline
CLM + MCQA	CLM	**+7.1pp**
MLM + MCQA	MLM	−2.8pp

Finding: MLM *degrades* with instruction data

→ MLM's sparse gradients can't exploit structured Q&A

→ CLM's uniform coverage captures the signal

Ch.5: Position Trade-off

Position	CLM	MLM	Δ
256–1024	0.918	0.873	**+4.5%**
1024–2048	0.913	0.875	**+3.8%**
2048–4096	0.723	0.834	−11%
4096+	0.635	0.757	−12%

- CLM wins at 0–2048
- MLM wins at 2048+
- Most clinical docs: 256–2048

Why?

Gradient imbalance in CLM:

- Early positions: over-trained
- Late positions: under-trained

ModernCamemBERT-bio: CLM vs MLM (8 tasks, 9 seeds)

Type	Task	MCB-bio (MLM)	MCB-bio (CLM)	Δ
ICD-10	FRACCO-30	66.8	**71.0**	+4.3pp
ICD-10	FRACCO-100	54.4	**57.1**	+2.6pp
SNOMED	Cantemist	62.1	**64.6**	+2.5pp
SNOMED	Distemist	**28.1**	23.9	−4.2pp
Dialog	MedDialog	62.4	**63.8**	+1.4pp
Classif.	DiaMED	59.3	**64.1**	+4.8pp
NER	EMEA	69.1	**71.2**	+2.1pp
NER	Medline	59.8	**62.1**	+2.2pp
Average		57.7	**59.7**	**+2.0pp**

CLM wins **7/8 tasks** → robust across task types

Ch.5: English Results (PubMed 10B)

ModernBERT-base on English biomedical benchmarks

Type	Task	MLM	CLM	Δ
QA	PubMedQA	**47.8**	34.6	−13.2pp
MCQA	MedQA	17.6	**18.2**	+0.6pp
Classif.	GAD	70.6	**73.9**	+3.2pp
Relation	ChemProt	28.4	**29.9**	+1.5pp

CLM wins 3/4 tasks – similar pattern to French

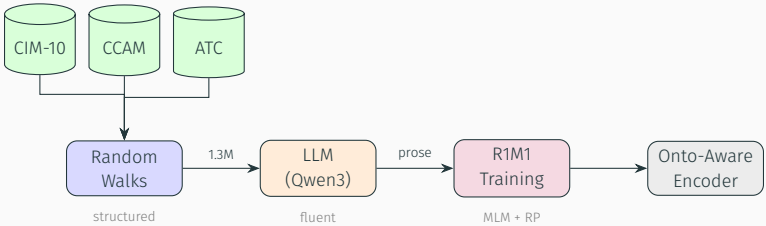
Exception: QA requires bidirectional attention (question↔context)

→ CLM's causal attention hurts Q↔A reasoning

Can we also inject structured knowledge

during pretraining?

Publication: Ontobook – Inject ontology knowledge during pretraining



Ontology	Walks	Size
CIM-10 (diagnoses)	402k	2.3 GB
CCAM (procedures)	763k	2.9 GB
ATC (drugs)	139k	291 MB

Transform structured walk $\rightarrow$ fluent medical prose

Structured Walk:

```
[E11] Diabète de type 2
>> Partie de: E10-E14
>> À distinguer de: E10
>> Complications: E11.2, E11.3
```

$\Rightarrow$

Fluent Textbook:

Le **diabète de type 2** (E11) appartient aux diabètes sucrés (E10-E14). Il doit être distingué du diabète de type 1 (E10). Ses complications incluent les atteintes rénales (E11.2) et ophtalmiques (E11.3).

Training objective:

- MLM: predict masked tokens
- Relation Prediction: 6 classes (parent, child, sibling, etc.)

Finding: +3.86 points over MLM baseline

Model	FRACCO	Cantemist	Distemist	Avg
MLM-baseline	55.81	66.01	24.23	48.68
OntoBook	**58.33**	**67.06**	**32.24**	**52.54**
Misaligned	33.39	20.11	12.63	22.04

Critical insights:

- **Alignment is essential:** misaligned = −26.64 points (catastrophic)
- **Cross-ontology transfer:** ATC (drugs) improves Distemist (diseases) +8 pts

Part 3: Adapting to Clinical Tasks

How to use pretrained models with little annotation?

Ch.7: The Problem – Too Many Labels

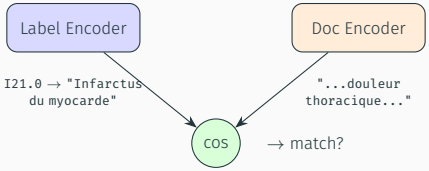
Clinical coding: thousands of labels

Task	Ontology	Labels
Diagnosis coding	ICD-10	17,000+ codes
Drug coding	ATC	6,000+ codes
Procedure coding	CCAM	8,000+ codes
Entity linking	SNOMED-CT	350,000+ concepts

Classic BERT NER:

- Train classifier head with N outputs (N = labels)
- Need annotated examples for *each* label
- $\rightarrow$ 17k labels = impossible to annotate

Publication: MCB-bio-embed (MCB-bio-gliner)



Key insight: No classifier head → zero-shot on *any* label set

But: Standard GLiNER doesn't exploit ontology structure!

Ch.7: WHY RDF? – The Opportunity

Problem: Standard GLiNER ignores ontology structure

RDF/SKOS gives us everything for free:

1. Precise code↔label:
`I21.0` → `"Infarctus aigu"`

2. Synonyms (positives):
`I21.0 = "IDM antérieur"`

3. Hierarchy (positives):
`I21.0` ∈ `I21 (Infarctus)`

4. Exclusions = hard negatives!
`I21` ≠ `I25 "ancien infarctus"`
"à ne pas confondre avec..."

5. Related = hard negatives!
`I21` ≠ `I22 "récidive"`
"exclut le code..."

→ Perfect for sentence transformer training:

- 258k pairs with natural hard negatives
- Train specialized **label encoder** (MCB-bio-embed)
- Keep MCB-bio for **span prediction**

Multi-source: 1.5M pairs

Source	Pairs
RDF ontologies	258k
Synthetic passages	490k
Persona queries	760k
Total	**1.5M**

Persona example:

Clinicien:
“Quelle prise en charge pour une tumeur trophoblastique?”

Patient:
“C’est quoi le choriocarcinome? C’est grave?”

Codeur:
“Code CIM-10 pour thrombopénie à la rifampicine?”

→ 5 personas for query diversity + hard negative mining via Solon

Good on biomedical, but worse on clinical.

How to improve clinical performance without clinical data?

Ch.8: Synthetic Clinical Data for GLiNER

Supervision: M2 intern Anh Thu Vu (AP-HP collaboration)

Problem: Need clinical training data, but it's private

Solution: LLM generates synthetic clinical reports + annotations

Prompt components:

- Admin info (age, sex, dates)
- ICD-10 codes (DP, DAS)
- Tumor info (TNM, biomarkers)
- NCCN treatment guidelines
- Note template (AP-HP style)

Generated dataset:

- 1,000 annotated reports
- 46,776 entity mentions
- 1,554 unique entity types
- Median length: 1,429 tokens

Ch.8: Synthetic Report Example

Prompt:

```
Âge: 61, Sexe: M, Nom: L. Cornuche
Service: CANCERO, Hôpital: Lyon Sud
DP: Tumeur poumon (C349)
DAS: Hémiplégie (G819)
Stade: IV, EGFR+, ALK-, PD-L1+
Protocole: Pemetrexed+Cisplatine
```

⇒

Generated note:

"M. Cornuche, 61 ans, HDJ pour C3 chimiothérapie. ATCD: adénocarcinome pulmonaire stade IV, métastases cérébrales, hémiplégie séquellaire. Ttt: Pemetrexed-Cisplatine. Évolution favorable, sortie domicile..."

Extracted entities:

[MALADIE] adénocarcinome pulmonaire · [STADE] stade IV · [TRAITEMENT] Pemetrexed-Cisplatine · [METASTASE] métastases cérébrales

Blind evaluation: 50 synthetic vs 50 real notes (10-point scale)

	Language	Coherence	Completeness	Synthesis	Overall
Mistral AI	**9.33**	7.93	**9.67**	8.58	8.40
Real (AP-HP)	9.20	**9.33**	9.24	**9.02**	**8.69**

Key findings:

- Synthetic data rated **more complete** than real notes (9.67 vs 9.24)
- Real notes have better **medical coherence** (9.33 vs 7.93)
- Overall quality nearly identical (Δ = 0.29)

Ch.8: GLiNER Training Results

6 models trained on synthetic data

Model	Params	F1
GLiNER-Small	166M	62.05
GLiNER-Medium	209M	63.03
GLiNER-Multi	209M	63.49
GLiNER-Large	459M	**64.29**
GLiNER-CamemBERT-bio	135M	60.04
GLiNER-ModernBERT	173M	46.58

Zero-shot capability confirmed:

- ICD codes with $<$10 occurrences: F1 = 0.61
- 6/44 zero-shot entity types: F1 = 1.0 (perfect)
- 25/44 zero-shot entities: F1 $>$ 0.5

Can we combine ontology structure with clinical training?

Ch.9: Ontology-Aware Clinical NER

Combine insights from Ch.7 + Ch.8:

- MCB-bio-embed label encoder (RDF-trained)
- Synthetic clinical data for span encoder
- End-to-end GLiNER with both components

Hypothesis:

- RDF structure → better label embeddings
- Synthetic clinical data → better span detection
- Combined → best of both worlds

Contributions: Corpora & Models

Corpora:

- biomed-fr: 413M words (French biomedical)
- BE-Enriched: 2M clinical paragraphs from PMC
- GAPeron: 3.1T tokens French web (quality-filtered)

Models:

- CamemBERT-bio, ModernCamemBERT-bio (French encoders)
- BioClinical-ModernBERT (English SOTA encoder)
- GAPeron 1.5B / 8B / 24B (French LLMs)
- MCB-bio-embed (French biomedical bi-encoder)

Contributions: Methods

Methods:

- BiaHS: Benchmark-in-a-Haystack (contamination detection)
- CLM→decay: computational hysteresis for encoder CPT
- OntoBook: ontology-to-textbook reformulation
- RDF-based sentence transformer training (MCB-bio-embed)

Contributions: Libraries

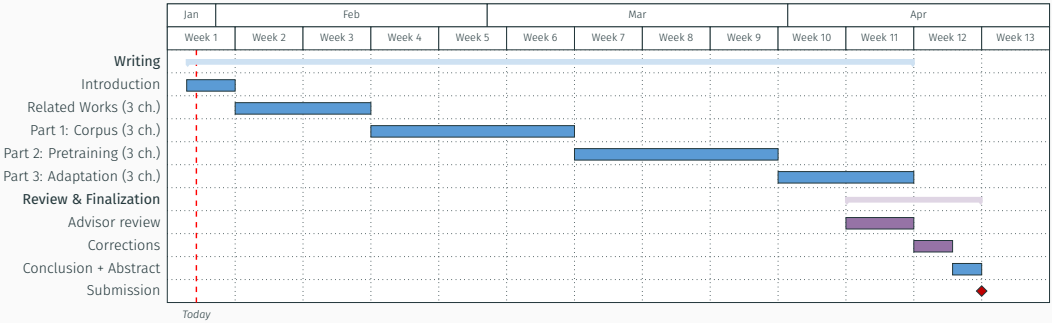
TABIB – Tokenizer-Agnostic Biomedical Information extraction Benchmark

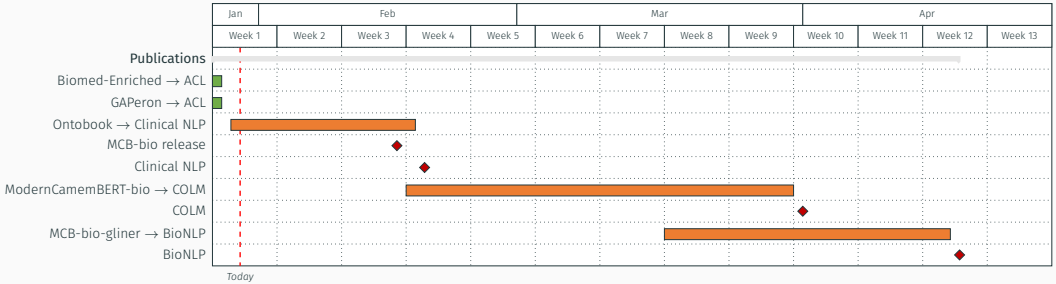
- Evaluate LLM vs BERT vs GLiNER fairly
- Evaluation NOT influenced by tokenization differences
- Addresses issue from CamemBERT-bio: token-level eval favors certain tokenizers

Chainette – Structured LLM chains with vLLM

- Pydantic typing for structured outputs
- Branches, nodes, routers for complex pipelines
- Used for: IE, reformulation, synthetic data generation

Timeline


Jan
Feb
Mar
Apr
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Week 9
Week 10
Week 11
Week 12
Week 13
Writing
Introduction
Related Works (3 ch.)
Part 1: Corpus (3 ch.)
Part 2: Pretraining (3 ch.)
Part 3: Adaptation (3 ch.)
Review & Finalization
Advisor review
Corrections
Conclusion + Abstract
Submission
Today
Writing
Review


Jan
Feb
Mar
Apr
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Week 9
Week 10
Week 11
Week 12
Week 13
Publications
Biomed-Enriched → ACL
GAPeron → ACL
Ontobook → Clinical NLP
MCB-bio release
Clinical NLP
ModernCamemBERT-bio → COLM
COLM
MCB-bio-gliner → BioNLP
BioNLP
Today
Submitted
To do