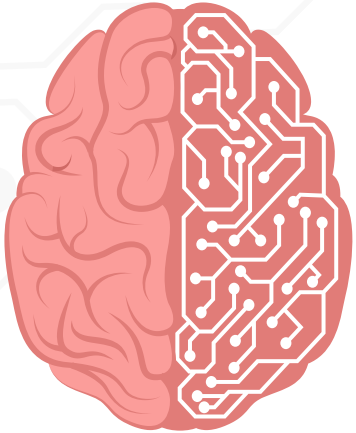


FullMouth

Using LLMs for Enhancing Quality of Dental Clinical Data

Bunmi Tokede

University of Texas Health Science Center at Houston

IADR/AADOCR/CADR General Session | San Diego, CA | March 2026

Overview

01

The Data Quality Challenge

02

Why LLMs Change the Equation

03

The FullMouth Approach

04

Results & Insights

05

Implications & Next Steps

BigMouth

Dental Data Repository

7M

Patient Records

19

Partner Institutions

26

Years of Data

PARTNER INSTITUTIONS

- Harvard
- UTHealth Houston
- U of Michigan
- Loma Linda
- Texas Tech
- U of Puerto Rico
- Stony Brook
- UCSF
- U of Iowa
- UC Denver
- U of Buffalo
- U of Minnesota
- Methodist Healthcare
- Tufts
- U of Pittsburgh
- U of Florida
- Marquette
- U of Pennsylvania
- VCU

Largest clinical dental data repository in the US · **Hosted at UTHealth Houston** · **Initial funding by NLM**

Data Currently in BigMouth

Six interconnected data domains capturing the full patient record

D

Demographics

Age, sex, zip codes, language, etc.

M

Medical

Medical history, medications, surgeries

O

Odontogram

Teeth present, missing, decayed, restored, etc.

I

Insurance

How patients pay for services

De

Dental

Dental history, diagnoses, procedures

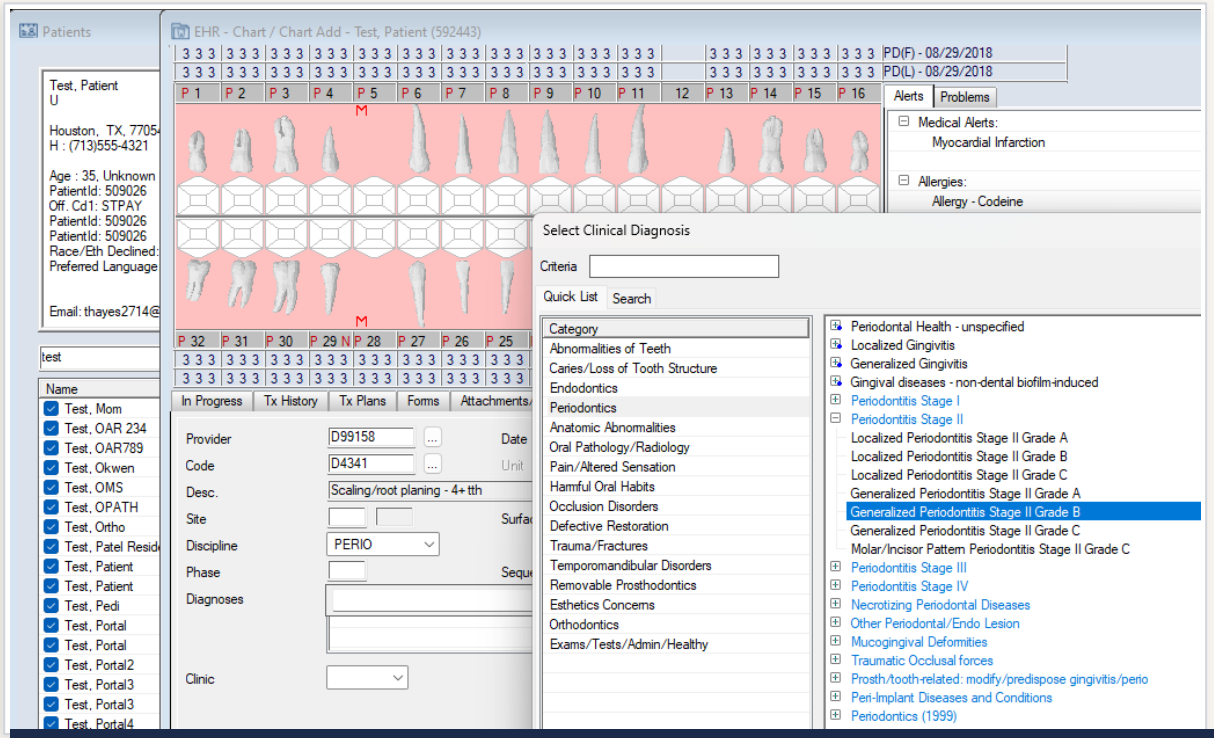
P

Periodontal

CAL, pocket depth, bleeding on probing, etc.

Structured data: what we are referring to

EHR systems capture clinical information through coded fields — but coverage is incomplete



Odontogram

Visual tooth chart with coded conditions (present, missing, decayed, restored)

Coded Diagnoses

ICD, SNODDS etc.

Procedure Codes

CDT codes (e.g., D4341)

Medical Alerts

E.g., Allergies

Advantage: Clean, queryable data **Challenge:** Rigid templates, incomplete adoption — many fields left empty

BigMouth Data Completeness

Percentage of records with structured data populated

- These 4 elements are representative.
- Many other clinical data fields show similar or worse patterns across BigMouth sites.

Previous Strategies — Limited Impact

Forcing Functions

Mandatory fields in EHR workflows to require data entry before proceeding.

Providers find workarounds; data quality doesn't meaningfully improve.

Checklists

Standardized checklists for clinicians to ensure completeness at point of care.

Added burden on clinicians; compliance inconsistent across providers.

The Information is There — It's Just Unstructured

Sample Clinical Progress Note (de-identified)

78 yo Hispanic female pt presents to clinic
D- Localized Periodontitis Stage II Grade A, pt presents for perio re-evaluation after SRP.
H- Med hx reviewed. No changes. Pt reports finishing antibiotics. Pt is continually taking medication for HTN, seizures, and liver cirrhosis. ASA III.
O- Patient states she brushes 2X per day, using floss for interdental hygiene. Oral hygiene is fair. Plaque present throughout the dentition...
Bleeding Scores: Initial BOP 25%, at re-eval BOP was 33%.
Mobility changes: No mobility was found at the initial exam or re-evaluation exam.

02

Why LLMs Change the Equation

From task-specific models to general-purpose language understanding

What Already Works in Clinical NLP

Strong results are possible on narrow, well-defined extraction tasks

What works

- BERT-family models (BioBERT, ClinicalBERT, RoBERTa) are the established standard for clinical NER.
- Strong per-task performance when fine-tuned on domain-specific annotated data.

OUR PRIOR WORK

Chuang, Tokede et al. (AMIA 2023)

RoBERTa for perio Dx → F1 0.92–0.97

Tokede et al. (JAMIA Open 2025)

GPT-4 synthetic notes + RoBERTa → cross-site 0.98–0.99

These approaches work — for single-entity extraction tasks

Why We Moved Beyond RoBERTa

Two practical arguments for the LLM approach

01

Multi-entity extraction at scale

Under the traditional approach, multi-entity extraction means:

1. multiple separate annotation efforts,
2. multiple models,
3. multiple maintenance pipelines — none of which transfer across sites.

02

Annotation efficiency

GPT-4 achieves F1 0.60–0.80 through prompt engineering alone (Hu et al. JAMIA 2024).

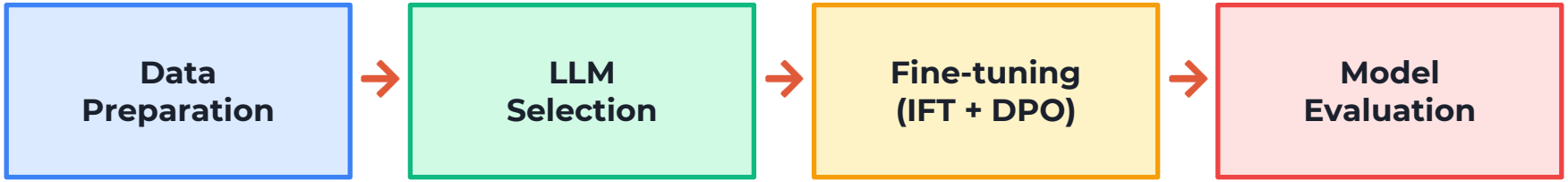
Fine-tuning with modest datasets (~400–500 examples) reaches diminishing returns — a fraction of the annotation burden required by encoder models.

03

The FullMouth Approach

Data preparation, model selection, and fine-tuning for dental NER

FullMouth Project Overview



26,085 notes

8,905 patients
CY 2023

Multiple LLM families

Open-source

3-stage pipeline

Prompt → IFT → DPO

19 entity types

Precision, Recall
F1-score

Study Design Highlights

Inclusion: Progress notes from exams (D0150, D0120, D0180) and perio procedures (D1110, D4341, D4342, D4910)

Target Entities: 19 Clinical Data Points

Demographics

Age, Race, Ethnicity, Sex

Periodontal Dx

Diagnosis, Stage, Grade, Extent, Subtype

Medical History

Systemic Condition, Family Hx, Previous Procedure

Medications

Medication Allergy, Medication Taken

Lab / Social

HbA1c Levels, Social Factors

Home Care

Brushing freq., Flossing, Other Home Care

Annotations results:

Age

- '74'

Sex

- 'F'

Systemic Condition

- 'Hypertension'
- 'High cholesterol'
- 'arthritis'

Previous Medical Procedure

- 'C-section (1990, 1988)'
- 'gallbladder removal (2015)'

Medication Allergy

- 'Cefazulin'
- 'Vancomycin'
- 'Nitrofurantoin'

Medication Taken

- 'Meloxicam'

Enter Label Name | Add Label | Age | Race | Ethnicity | Sex | Perio Diagnoses | Stage | Grade | Extent | Subtype | Social Factors | Systemic Condition | Family History Disease | Previous Medical Procedure | Medication Allergy | Medication Taken | HbA1c Levels | Brushing freq | Flossing | Other Home Care

Grad Prosth Comp Exam -
- -
- D: -
- 74 F presents to grad for comprehensive exam -
- -
- H: -
- MedHx form reviewed and updated -
- PMHx: Hypertension, High cholesterol -
- PSHx: C-section (1990, 1988), gallbladder removal (2015) -
- SocHx: Alcohol (-), tobacco (-) , illicits (-) -
- Meds: Meloxicam (arthritis) (15 mg/day), Hydrochlorothiazide (25 mg), Losartan (100 mg). Solifenacin (5 mg), Pyridostigmine (60 mg), Ozempic (0.25 mg), Amlodipine (5-20 mg), Tylenol (325 mg), Melatonin, Vit D3, omeprazole (20 mg), Metoclopramide HCL (5mg/5ml), rosuvastatin (5mg) -
- Allergies: Cefazulin, Vancomycin, Nitrofurantoin -
- ASA: II -
- -
- -
- BP: 117/67 HR: 77 -
- O: Oral Hygiene: Poor -
- T: -
- EO exam completed visually and by bilateral palpation. ROM, MM, TMJ, and LAD all WNL. -
- IO exam conducted visually and with bilateral palpation of buccal/labial mucosa, dorsal/lateral/ventral surfaces of tongue, FOM, pharynx, and hard/soft palate. WNL -
- Caries risk assessment: (High) -
- Hard tissue charting: completed and noted -
- Dx cast taken using Trios. -
- Oral Cancer Screening Negative. -

Annotation & Evaluation Dataset

950

annotated progress notes

Calibration

1st	50 notes	IRR —	*Defined guidelines*
2nd	50 notes	IRR 51.8%	*Refined guidelines*
3rd	50 notes	IRR 77.5%	*Finalized guidelines*
4th	50 notes	IRR 89.7%	

200

for fine-tuning

750

gold standard for evaluation

Model Selection

Open LLM Leaderboard · filtered by deployability and instruction-following

Parameters: 1 – 25 B

Must run on institutional GPU hardware without sending PHI off-site.

Ranked by IFEval

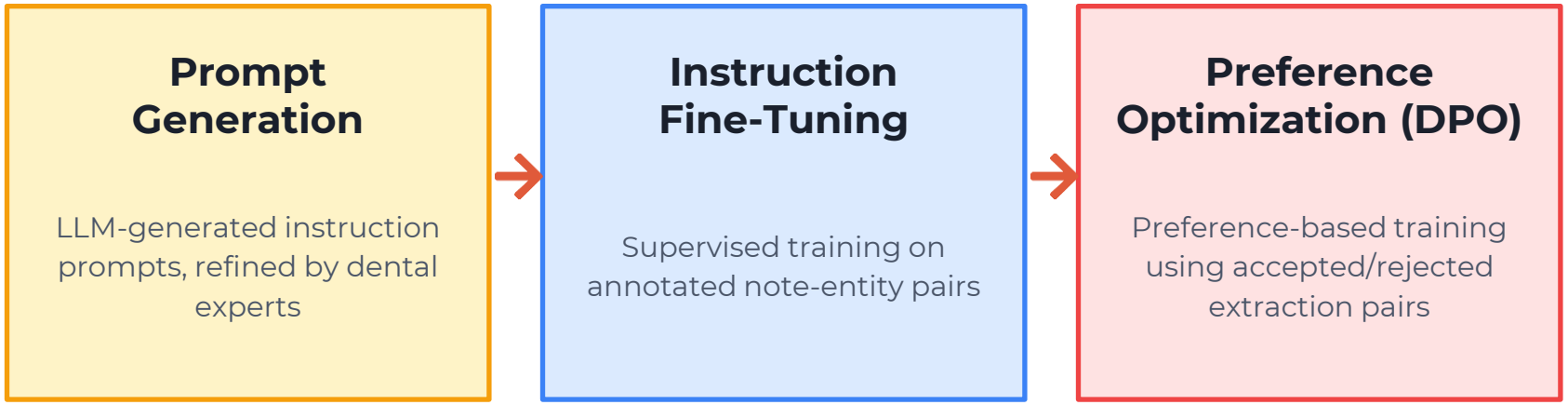
Instruction-following is the core capability for structured extraction prompts.

10 models included across 6 families

Model	Params
Qwen2.5-14B-Instruct	14B
Qwen2.5-7B-Instruct	7B
Llama-3.1-8B-Instruct	8B
Llama-3.2-3B-Instruct	3B
OLMo-2-13B-DPO	13B

Model	Params
OLMo-2-7B-DPO	7B
Gemma-2-9B-it	9B
Gemma-3-12B-it	12B
Mistral-Small-2409	22B
Granite-3.1-8B-Instruct	8B

Fine-Tuning Pipeline: IFT + DPO



Prompt Engineering: Three Strategies

Progressively richer instructions — no model weight changes

STRATEGY 01

Definitions Only

Entity type definitions provided to the model with no examples or guidance.

EXAMPLE

"RACE: self-identified or documented racial category"

F1: 0.625

→

STRATEGY 02

+ Examples

Added concrete examples for each entity type to guide extraction behavior.

EXAMPLE

"RACE: White, African American, Asian"

F1: 0.652

→

STRATEGY 03

+ Error Feedback

Targeted corrections based on observed model mistakes across entity types.

EXAMPLE

"Race ≠ Ethnicity — label 'White' as RACE"

F1: 0.680

Prompt design yielded steady gains *— but could only go so far. Fine-tuning was needed to go further.*

FINE-TUNING

Instruction Tuning

INSTRUCTION

You are doing clinical **NER** over de-identified notes.
Extract spans for: DIAGNOSIS, RACE_ETHNICITY, ALLERGY_MED

INPUT

Note: Mr. K is an **African American** male with uncontrolled **hypertension**.
He has a documented anaphylactic reaction to **ceftriaxone**.

OUTPUT

{"text": **"African American"**, "label": **"RACE"**}

{"text": **"hypertension"**, "label": **"DIAGNOSIS"**}

{"text": **"ceftriaxone"**, "label": **"ALLERGY_MED"**}

DPO

Direct Preference Optimization

Option A (preferred)

{"text": **"Hispanic"**, "label": **"ETHNICITY"**}

{"text": **"type 2 diabetes"**, "label": **"MED_HISTORY"**}

{"text": **"hypertension"**, "label": **" MED_HISTORY"**}

{"text": **"amoxicillin"**, "label": **"ALLERGY_MED"**}

Option B (rejected)

{"text": **"poorly controlled type 2 diabetes"**, "label": **" MED_HISTORY"**}

// span too long

{"text": **"lip swelling"**, "label": **" MED_HISTORY"**}

// symptom, not a diagnosis

{"text": **"asthma"**, "label": **" MED_HISTORY"**}

// note says "No history of asthma"

04

Results & Insights

What worked, what's hard, and what it means

RESULTS

Model performance across training

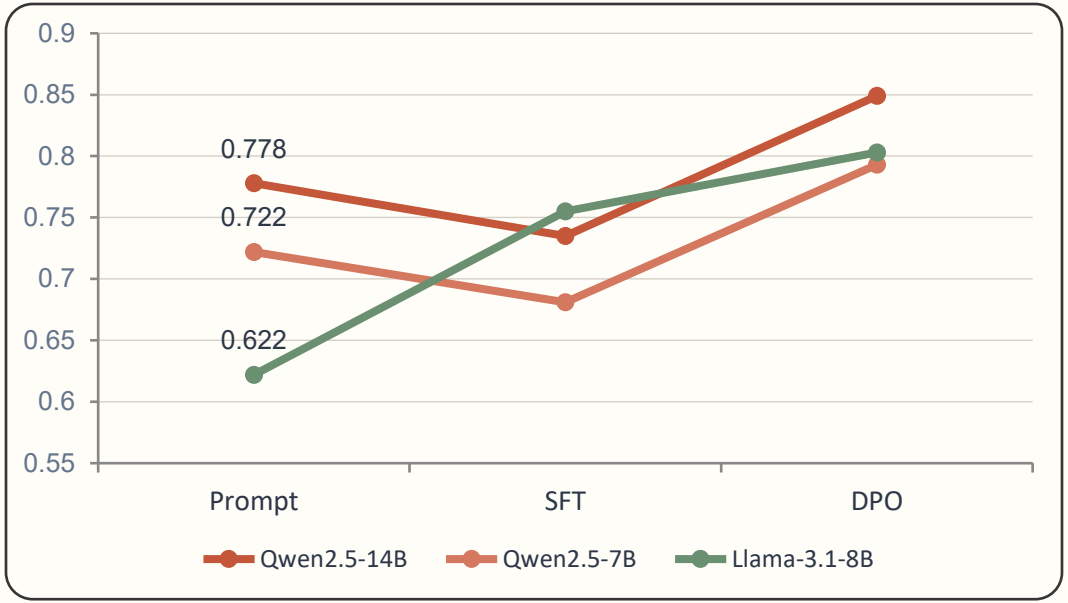
Model	Size	Prompt	SFT	Δ SFT
Qwen2.5-Instruct	14B	0.778	0.735	**-0.043**
Qwen2.5-Instruct	7B	0.722	0.681	**-0.041**
Granite-3.1	8B	0.644	0.534	**-0.110**
OLMo-2-DPO	13B	0.641	0.574	**-0.067**
Gemma-2-it	9B	0.627	0.598	**-0.029**
Llama-3.1-Instruct	8B	0.622	0.755	**+0.133**
OLMo-2-DPO	7B	0.419	0.658	**+0.239**
Mistral-Small	22B	0.375	0.494	**+0.119**
Llama-3.2-Instruct	3B	0.336	0.510	**+0.174**

reduced performance

improved performance

The DPO Effect

A non-linear training trajectory



What's happening

For Qwen, DPO recovers and surpasses prompt-only performance.

Llama shows a different pattern: steady, linear improvement through all stages.

The path to best performance is not always linear. DPO's preference alignment can rescue models that SFT destabilized

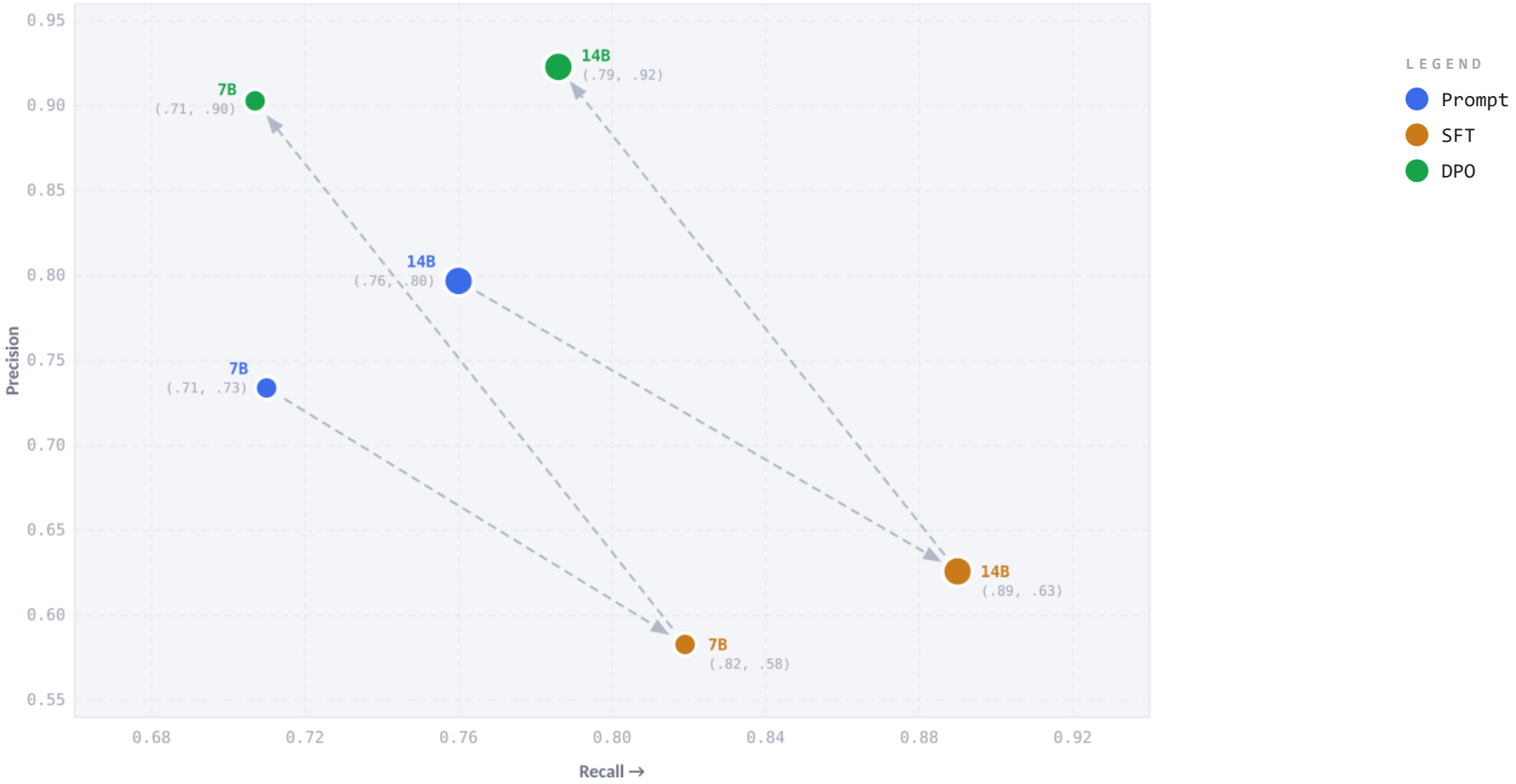
Precision vs. Recall

Qwen2.5-14B-Instruct + DPO

F1	Precision	Recall
0.849	**0.923**	**0.786**

A false positive contaminates the research database

Precision–Recall Tradeoff Across Training



Entity-Level Performance

Qwen2.5-14B-Instruct + DPO

Strong F1 ≥ 0.80

Ethnicity **0.97**
Stage **0.96**
Brushing Freq. **0.96**
Extent **0.92**
HbA1c Levels **0.91**
Grade **0.89**
Perio Dx **0.89**
Med Taken **0.87**
Subtype **0.87**
Sex **0.86**
Med Allergy **0.86**
Flossing **0.85**
Systemic Cond. **0.85**

Moderate F1 0.60–0.80

Family Hx 0.74
Social Factors 0.74
Age 0.73
Prev. Procedure 0.70
Race 0.60

Challenging F1 < 0.60

Other Home Care 0.31

05

Implications & Next Steps

From research prototype to clinical data infrastructure

What the Results Show

1

Prompt-only evaluation can help screen models before fine-tuning

Zero- and few- shot prompting provides a low-cost way to evaluate LLM candidates before investing in fine-tuning

2

Smaller models can be competitive for clinical NER

Model size alone does not predict NER performance. Llama-3.1 8B (0.8 F1) outperforms much larger models after fine-tuning, opening the door to efficient, deployable solutions.

3

With modest annotation, LLMs had strong performance

With a modest set of annotated clinical notes, LLMs fine-tuned via SFT produce good results while offering greater flexibility across entity types.

Next Steps

01

Error Analysis

Systematic investigation of failure modes across entity types to guide targeted improvements.

02

Multi-Site Generalizability

Test extraction quality across BigMouth partner institutions with different documentation styles.

03

Computational Cost & Scalability

Benchmark inference cost per note, GPU-hours for processing, and cost–performance tradeoffs across model sizes.

04

Model Finalization & Deployment

Move from research prototype to a production pipeline integrated into BigMouth data infrastructure.

Long-term vision: A validated, deployable NLP pipeline that transforms unstructured dental notes into research-ready structured data across the BigMouth network.

Thank You

oluwabunmi.tokede@uth.tmc.edu

Supported by the National Institute of Dental and Craniofacial Research: R56DE034086