

SerialSentence: A Date-Centric OMOP-CDM Reorganization for LLM Reasoning

Yoon Seob Kim^{1,2,3}, Don Gueu Park⁴, Jung Han Yoon⁴, Chungsoo Kim^{1,2,3}, Rae Woong Park^{1,2,3,5}

¹*Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea*

²*Center for Biomedical Informatics Research, Ajou University Medical Center, Suwon, Republic of Korea*

³*Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea*

⁴*Department of Neurology, Ajou University School of Medicine, Suwon, Republic of Korea*

⁵*BK21 R&E Initiative for Advanced Precision Medicine, Suwon, Republic of Korea*

Background

Large language models (LLMs) can reason over a patient's entire clinical history, which shows promise for reduced documentation work burden and benefits in informed clinical decision making.^{1,2} Translating electronic medical records (EMR) into LLM-ready input can be a difficult task due to limited context window of language models and the continuously growing size of records. Reconstructing encounters only from structured tabular data loses patient-specific text records, while utilizing solely the clinical notes omits valuable information such as exact lab values, drug doses, etc. For investigators who aim to apply LLMs to EMR data in entirety, a serialization that is robust to integrate both structured data and free text records is needed. We present SerialSentence, a date-centric structure that converts the full warehouse into per-patient timelines designed for whole-history LLM reasoning, with a guaranteed-fit context strategy that scales single-pass LLM summarization to a 184,090-patient neurology cohort.

Methods

Data Sources

We used the Ajou University School of Medicine (AUSOM) database, a Korean tertiary-hospital OMOP-CDM with ≥ 2.9 million patients (1994–2021) and a NOTE table of clinical free text. SerialSentence was built for the 184,090 patients with at least one neurology visit (1994–2025). Diagnosis recovery was validated against a neurologist-adjudicated gold-standard table of hand-reviewed movement-disorder patients; the analysis population is the 1,526 patients whose adjudicated diagnosis is one of the eight parkinsonian syndromes or essential tremor expressible as an OMOP condition code (Table 1); clinically uncertain cases were excluded.

The SerialSentence Structure

SerialSentence is a data structure where each date on which any event occurs across eight OMOP tables (CONDITION_OCCURRENCE, DRUG_EXPOSURE, MEASUREMENT, PROCEDURE_OCCURRENCE, OBSERVATION, DEVICE_EXPOSURE, NOTE, and VISIT metadata) emits one structured JSON in sentence format readily recognizable by LLM. (Figure 1) For example, drug status is derived from overlapping DRUG_EXPOSURE intervals into transitions (started / active / stopped), days-supply, and dose unit, and all concept_id fields are resolved into plain concept names. The build is two-stage: stage 1 emits a concept-resolved (concept codes translated to plain language) daily base layer; stage 2 selects the notes that enter the model's context window before LLM input.

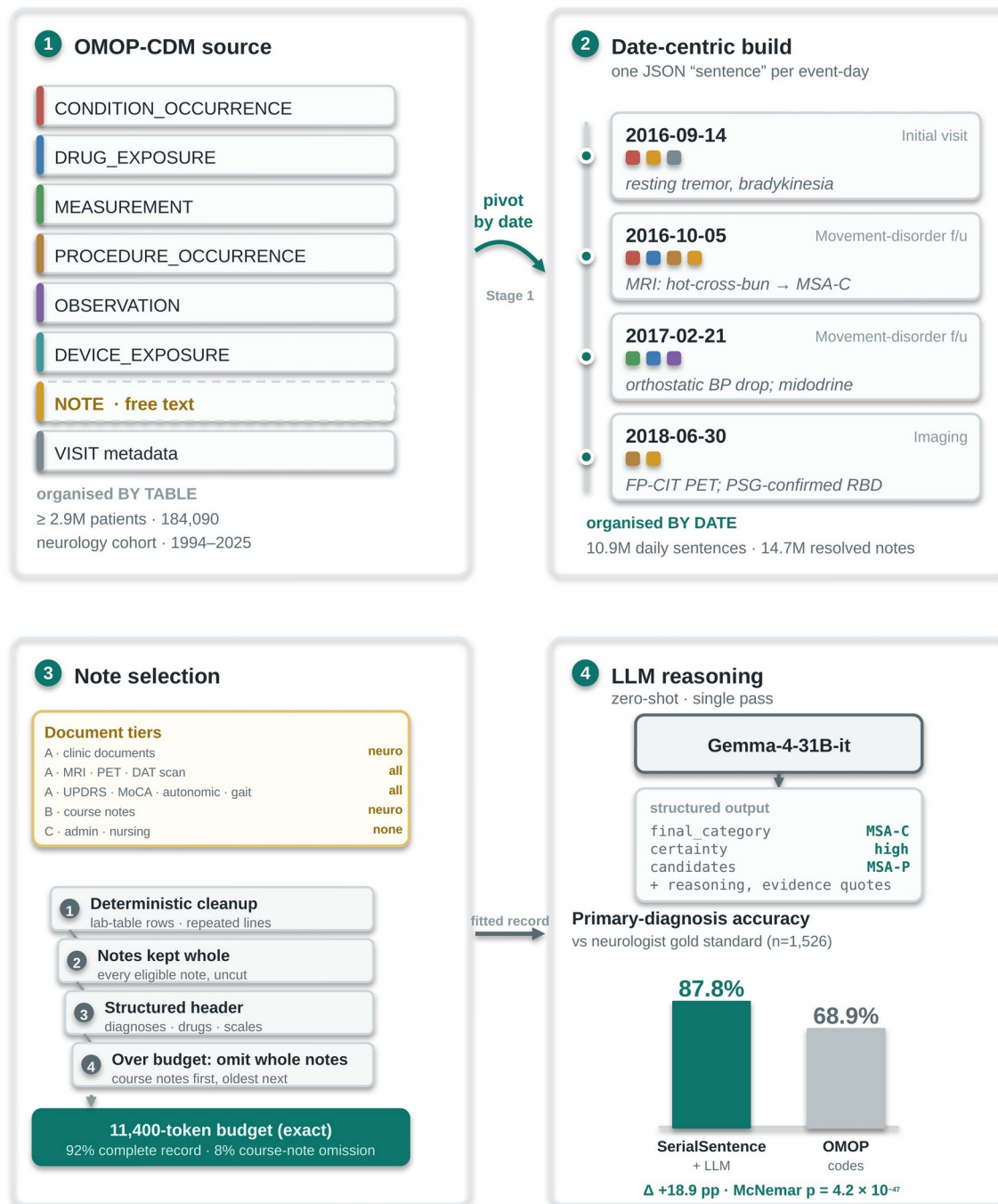


Figure 1. The SerialSentence Framework

Note selection and context fitting

Notes enter the prompt whole — none is truncated. Each note’s authoring department is resolved

through OMOP PROVIDER and CARE_SITE. Initial-visit, admission and discharge documents (tier A) and clinical-course notes (tier B) are included only when authored by Neurology or Neurosurgery; imaging and neurological test reports are included from every department (tier A); administrative and nursing documentation is excluded (tier C). Two deterministic cleanup rules remove copy-pasted laboratory-result tables and exact repeats of earlier lines within a note. A structured header summarizing other OMOP tables (coded neurological diagnoses, drug timeline, rating scales, procedures) precedes the notes. The prompt is capped at 11,400 Gemma-4 tokens (set by two 40 GB A100 GPUs); when exceeded, whole documents are omitted by a predetermined rule: tier B notes are thinned by even subsampling of the chronological sequence, and only if tier A alone exceeds the limit are its oldest documents dropped, always keeping the first initial visit.

Diagnoses were extracted with Gemma-4-31B-it (zero-shot, temperature 0.1) in one call per patient, returning JSON with the diagnostic category, a self-rated certainty and supporting evidence; an explicit 'Uncertain' category allows abstention. For reporting, certainty is mapped to LLM-certainty-high, -medium and -low (abstention). Errors are decomposed into incorrect diagnoses (a false positive for the asserted category and a false negative for the adjudicated one) and abstentions (false negatives only); per-category false positives and negatives are summarized by precision and recall (Table 1). The OMOP baseline assigns the most specific condition code from a curated SNOMED concept set with CONCEPT_ANCESTOR descendants. Accuracy, Cohen's κ and macro-F1 are compared on identical patients with bootstrap 95% CIs and an exact McNemar test.

Results

The complete build produced 10.9M daily sentences and 14.7M resolved clinical notes across 184,090 patients. Against the 11,400-token prompt budget, 91.9% of patients' complete eligible records were included whole; 8.1% required omission of some course notes and 2.1% omission of older diagnostic documents. No note was truncated.

On the 1,526 patients, the LLM achieved 87.8% primary-diagnosis accuracy versus 68.9% for OMOP codes (paired Δ +18.9 pp [16.4–21.6], McNemar $p=4.2\times 10^{-47}$), κ 0.813 vs 0.507 and macro-F1 0.831 vs 0.555 (Table 1). Table 2 stratifies accuracy by self-rated certainty: LLM-certainty-high covered 72.9% of patients at 96.2% (macro-F1 0.920), LLM-certainty-medium 24.7% at 71.4% (0.724), and the model abstained on 2.4%. Overall the LLM asserted an incorrect diagnosis for 9.8% of patients (false positives) and failed to return the adjudicated diagnosis for 12.2% (false negatives, including 2.4% abstentions), versus 18.0% and 31.1% for OMOP codes.

Table 1. Diagnosis recovery on the neurologist-adjudicated gold standard (n=1,526): zero-shot Gemma-4-31B on SerialSentence versus OMOP condition codes (precision / recall / F1 per diagnosis).

Diagnosis	N	Gemma-4-31B P / R / F1	OMOP codes P / R / F1	Δ F1
Parkinson's disease (PD)	884	0.97 / 0.92 / 0.94	0.85 / 0.90 / 0.87	+0.07
Essential tremor (ET)	147	0.96 / 0.83 / 0.89	0.69 / 0.20 / 0.31	+0.58
Drug-induced parkinsonism (DIP)	145	0.94 / 0.84 / 0.89	0.70 / 0.36 / 0.47	+0.41
Multiple system atrophy (MSA)	107	0.92 / 0.90 / 0.91	0.88 / 0.72 / 0.79	+0.12
Progressive supranuclear palsy (PSP)	99	0.79 / 0.79 / 0.79	0.93 / 0.55 / 0.69	+0.10
Dementia with Lewy bodies (DLB)	72	0.68 / 0.76 / 0.72	0.64 / 0.25 / 0.36	+0.36
Vascular parkinsonism (VP)	48	0.77 / 0.75 / 0.76	0.71 / 0.46 / 0.56	+0.20
Corticobasal syndrome (CBS)	24	0.81 / 0.71 / 0.76	0.86 / 0.25 / 0.39	+0.37
Overall (n=1,526)		macro-F1 0.831	macro-F1 0.555	+0.28

Table 2. Accuracy stratified by the model's self-rated certainty (n=1,526). An incorrect diagnosis (a committed category other than the adjudicated one) is a false positive for the asserted category and a false negative for the adjudicated one; an abstention ('Uncertain' for the LLM, no condition code for OMOP) is a false negative only. Any false negative = incorrect + abstained.

LLM certainty	Patients, n (%)	Accuracy / macro-F1	Incorrect diagnosis (FP + FN), n (%)	Abstained / uncoded (FN only), n (%)	Any false negative, n (%)
LLM predictions, all	1,526 (100.0%)	87.8% / 0.831	150 (9.8%)	36 (2.4%)	186 (12.2%)
LLM-certainty-high	1,113 (72.9%)	96.2% / 0.920	42 (3.8%)	0 (0.0%)	42 (3.8%)
LLM-certainty-medium	377 (24.7%)	71.4% / 0.724	108 (28.6%)	0 (0.0%)	108 (28.6%)
LLM-certainty-low (abstained)	36 (2.4%)	0.0% / 0.000	0 (0.0%)	36 (100.0%)	36 (100.0%)
OMOP condition codes	1,526 (100.0%)	68.9% / 0.555	274 (18.0%)	201 (13.2%)	475 (31.1%)

The gap reflects under-capture in coding and ETL practice rather than absence from the vocabulary: codes are precise but under-recalling (PSP precision 0.93 at recall 0.55; essential

tremor recall 0.20), while the two approaches are near-equivalent on Parkinson’s disease (F1 0.943 vs 0.874). The pipeline is thus complementary to coding, backfilling CONDITION_OCCURRENCE coverage under human review. Figure 2 shows an MSA-C case in which one zero-shot call reconstructs the diagnostic narrative (cerebellar predominance, hot-cross-bun sign, FP-CIT PET, levodopa non-response, PSG-confirmed RBD) — detail with no structured equivalent in the CDM.

Figure 2. Demonstration of a synthetic patient sample. (A) SerialSentence daily record (B) zero-shot LLM JSON report produced from the patient's fitted patient record.

(A) SerialSentence input (one date)	(B) LLM output JSON
<pre>{ "person_id": 9000001, "date": "2016-10-05", "day_index": 21, "visit": { "visit_type": "Outpatient Visit", "conditions": [{ "name": "Multiple system atrophy", "source_code": "G90.3" }], "drugs_active": [{ "name": "Midodrine", "status": "continuing", "quantity": 90.0, "days_supply": 30, "dose_unit": "TAB" }], "drugs_started": [], "procedures": [{ "name": "Brain MRI without contrast", "source_code": "70551" }], "notes": [{ "title": "Progress Note Movement Disorder", "type": "Follow-up", "text": "MRI: disproportional cerebellar atrophy; cross-line high signal " "in pons/medulla (hot-cross-bun sign).", "diagnosis": "MSA-C" }] } }</pre>	<pre>{ "cdm_id": 9000001, "extraction": { "reasoning": "Patient presented with a combination of cerebellar ataxia (dysarthria, dysmetria, truncal ataxia), parkinsonism (rigidity, bradykinesia), and severe autonomic failure (orthostatic hypotension, erectile dysfunction, and neurogenic detrusor underactivity requiring a Foley catheter). MRI showed disproportional cerebellar atrophy and 'cross line' high signal in the pons/medulla, which is highly characteristic of MSA-C.", "first_impression": { "primary": "MSA", "certainty": "probable" }, "final_diagnosis": { "primary": "MSA", "certainty": "confirmed", "subtype": { "motor_subtype": "MSA_C", "certainty": "clinically_established", "autonomic_severity": "severe", "cerebellar_features": "present", "stridor": "absent" } }, "comorbidities": [{ "condition": "Neurogenic Detrusor Underactivity", "relationship": "complication" }, { "condition": "Orthostatic Hypotension", </pre>

	<pre>"relationship": "concurrent"}, {"condition": "Cataract", "relationship": "concurrent"}] } }</pre>
--	---

Conclusions

SerialSentence bridges OMOP-CDM to whole-timeline LLM input that integrates clinical notes with the structured record. Preserving the full multimodal CDM retains signal that note-only pipelines discard, and whole-document fitting makes single-pass reasoning feasible within a fixed context budget. Zero-shot diagnosis recovery against neurologist adjudication unlocks latent clinical signal. The model’s self-rated certainty is a usable measure: the high-certainty stratum (72.9% of patients) was 96.2% accurate and can be assigned automatically, whereas the medium-or-lower stratum can be deferred to human review as capacity allows. Notably, the model errs by over-calling rather than withholding — an incorrect diagnosis for 9.8% of patients versus abstention for only 2.4%. Limitations: a single note-dense bilingual (Korean/English) institution and subspecialty; a process built for diagnosis retrieval only, so other extraction targets need their own; and the need for external replication. All computation used a locally served open-weights model, precluding need for protected health information leaving the site.

References

1. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* 2024;30:1134–42. doi:10.1038/s41591-024-02855-5
2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620:172–80. doi:10.1038/s41586-023-06291-2