ClinFuse: Patient Entity Resolution Powered by MedGemma Clinical Reasoning

Team

Alexander Rider — Solo submission.

Problem Statement

A cardiac arrest patient's record was confused with another patient's—one who carried a do-not-resuscitate order. The care team withheld life-saving treatment [1]. When healthcare systems can't reliably match records to the right patient, people die. Patients with duplicate charts—the most common form of this identity failure—are five times more likely to die during hospitalization [2]. Across 7,613 wrong-patient events at 181 healthcare organizations, 9% resulted in patient harm [1]. For the emergency department registrar matching an unconscious patient to their records, or the IT staff reconciling patient feeds from dozens of hospitals, these are not edge cases—they are daily realities.

Duplicate records aren't just dangerous—they're pervasive. An estimated 8–12% of hospital records are duplicates [3], costing the U.S. healthcare system over $6 billion annually [4]. Match accuracy averages 80–90% within institutions but falls to 50% in cross-organizational exchange [5]—meaning roughly half of cross-organizational duplicates go unresolved. The MATCH IT Act (H.R. 2002, 2025) sets a 99.9% matching target [6], and TEFCA v2.1 makes cross-organizational matching critical national infrastructure [7]. Closing the gap between these mandates and current cross-organizational accuracy requires a fundamentally different approach.

The barrier isn't better algorithms—it's better information. Any method relying exclusively on demographic fields faces an accuracy ceiling imposed by field quality—over half of duplicate pairs contain misspelled names or mismatched identifiers [3]. The missing signal is *clinical context*: the full breadth of a patient's medical record—conditions, medications, allergies, observations, and procedures—provides a powerful signal for resolving patient identity, regardless of demographic discrepancies. Yet no production entity resolution system exploits this signal: probabilistic and rule-based matchers lack the medical knowledge to interpret clinical histories as identity evidence. Resolving this limitation at scale could recover a substantial share of the $6 billion in annual duplication costs [4] and reduce the fivefold mortality risk [2] that fragmented records impose.

Overall Solution

Interpreting this clinical context as identity evidence requires medical reasoning—exactly the capability MedGemma [8] provides. To demonstrate its potential for patient entity resolution, we embed it in ClinFuse, a three-tier pipeline that augments fast probabilistic matching with MedGemma's clinical reasoning, invoked selectively for ambiguous cases. Figure 1 illustrates the architecture.

Tier 1: Blocking. All records enter Splink v4 [9], which first applies blocking rules on demographic fields to generate candidate pairs without exhaustive $O(n^2)$ comparisons.

Tier 2: Demographic Triage. Candidate pairs are scored via Fellegi–Sunter [10] probabilistic linkage using expectation-maximization with Jaro–Winkler string similarity and term-frequency adjustment. Pairs with high match probability are auto-matched; pairs with very low probability are auto-rejected. The remaining ambiguous "gray zone" pairs proceed to clinical review.

Tier 3: Clinical LLM. Gray-zone pairs receive structured clinical summaries and are classified by a fine-tuned MedGemma 4B model (described below). Each patient's clinical history is summarized into a compact format: conditions grouped by onset year, medications with date ranges, sorted allergies, latest values per vital sign, and procedures with years. The model sees two parallel summaries and classifies whether they describe the same patient (Figure 2). A Bayesian prior correction first shifts the LLM logit from the balanced training distribution to the gray zone's lower true-match rate, preventing overconfident match predictions. The corrected LLM logit and Splink match probability are then combined in log-odds space using interpretable linear weights.

Golden Records. Matched pairs form a graph whose connected components yield patient clusters. Field-level conflict resolution applies majority voting with domain heuristics (e.g., preferring longer address forms), producing a deduplicated Master Patient Index (MPI).

MedGemma as Clinical Reasoner. Prior LLM-based approaches address adjacent tasks—diagnosis-code linkage [11], clinical trial matching [12], general-domain entity resolution [13, 14, 15]—but none performs patient identity resolution using clinical narratives. Existing methods either rely on cloud-hosted models incompatible with HIPAA requirements, or operate exclusively on demographic fields [16] without leveraging clinical context.

MedGemma is a medically adapted model that combines clinical reasoning with edge-deployable size. Its medical fine-tuning embeds clinical knowledge that entity resolution in this domain demands: recognizing that "Essential Hypertension" and "Hypertension" denote the same diagnosis, that concurrent Hydrochlorothiazide and Lisinopril form a standard antihypertensive regimen, or that stable A1c trajectories suggest a single diabetic patient—knowledge a general-purpose model must acquire entirely from task-specific training data. Without this prior, a general-purpose model requires far more training examples and generalizes poorly to unseen terminology and clinical patterns. Figure 2 illustrates this reasoning on the evaluation set's hardest pair.

Technical Details

Data Generation and Augmentation

Synthea [17] generates realistic synthetic patients with clinically coherent histories spanning conditions, medications, allergies, observations, and procedures. An augmentation pipeline distributes each patient's records across multiple simulated facilities and

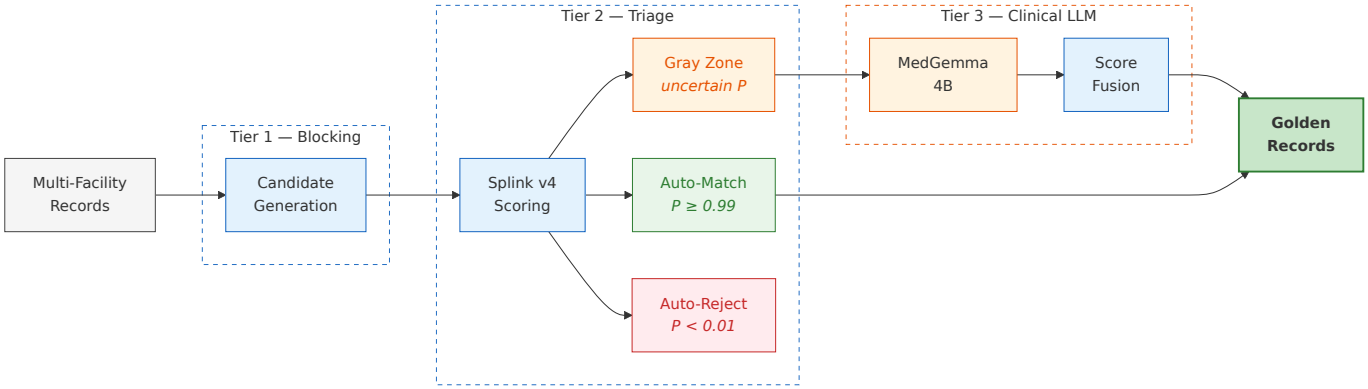

Figure 1: ClinFuse three-tier architecture. **Tier 1** generates candidate pairs via blocking rules. **Tier 2** scores pairs probabilistically; high-confidence matches (green) and clear non-matches (red) are resolved automatically. Only ambiguous gray-zone pairs (amber) route through the **Tier 3** clinical LLM for final resolution.

injects demographic errors: name variations (nickname substitution, typos, maiden name usage), address errors (abbreviation, format variation), date perturbations, identifier errors (SSN transposition, digit substitution), and formatting noise—mirroring documented real-world error patterns [3]. Separate datasets are generated for training (30,000 patients) and evaluation (2,500 patients, 6,264 records across 5 facilities).

Adapting MedGemma for Entity Resolution

Text-only adaptation. MedGemma 4B is a multimodal model with both vision and text encoders. Since our task is text-only, we strip the vision tower before fine-tuning, reducing from 4.2B to 3.88B parameters while preserving the medical language understanding from pretraining.

LoRA fine-tuning with classification head. We apply LoRA [18] (Low-Rank Adaptation) in bf16 precision targeting attention and MLP projection layers, and replace the generative language-model head with a two-class classification layer that outputs match/non-match logits directly—bypassing text generation entirely. Fine-tuned classification heads significantly outperform zero-shot prompting for binary tasks [19] while being orders of magnitude faster at inference [20]. Training uses balanced match/non-match pairs generated from augmented Synthea records.

Results

Pair-level metrics understate real-world impact, so Table 1 reports golden record quality—how many patients remain fragmented or unresolved—on 6,264 synthetic records (2,500 patients across 5 facilities) under deliberately adversarial conditions: each record carries 8 to 12 demographic errors and each patient's clinical history is split across a random number of facilities.

Table 1: Entity resolution on 6,264 records (2,500 patients, 5 facilities). Demographics Only uses Splink at its best F1 threshold. ClinFuse adds MedGemma gray-zone classification with score fusion.

Metric	Demographics	ClinFuse	Reduction
Split patients	421	**115**	73%
Fragmentation rate	14.4%	**3.9%**	73%
Unresolved patients	47.7%	**26.0%**	46%
Pair F1	0.764	**0.916**	—

Even at its optimal threshold, Splink leaves 421 patients with fragmented records—nearly one in six. MedGemma's clinical understanding—recognizing matching disease trajectories and overlapping medication regimens—reduces split patients by 73% to 115 and cuts the fragmentation rate from 14.4% to 3.9%, recovering 2,702 matches that demographics alone could not resolve. Pair-level F1 improves from 0.764 to 0.916, but the cluster-level impact is larger: each recovered link can transitively merge an entire patient cluster, so one correct LLM decision can unify records that would otherwise remain scattered across the MPI.

Projected impact. At an estimated 8–12% duplication rate [3] across roughly 36 million U.S. hospital discharges annually, approximately 3–4 million duplicate record pairs exist, implying a per-duplicate cost of ~$1,700 from the $6 billion aggregate [4]. ClinFuse's value concentrates in the cross-organizational gray zone, where matching accuracy is just 50% [5]. For a mid-size health information exchange processing 500,000 cross-organizational matches annually, roughly 250,000 pairs remain unresolved; if ClinFuse's 73% fragmentation reduction generalizes, that yields ~182,500 fewer split records—approximately $310M in recoverable costs for a single network, plus elimination of the associated fivefold mortality risk [2]. These projections extrapolate from synthetic evaluation; real-world impact depends on data quality, clinical history completeness, and institutional matching volumes.

Deployment and Feasibility

Hardware and privacy. MedGemma 4B runs in bf16 on a single consumer GPU (e.g., RTX 3090, ~$1,500) with all inference local—no patient data leaves the facility, supporting HIPAA/GDPR compliance and air-gapped deployment.

Cost at scale. The tiered architecture keeps GPU costs proportional to ambiguity: only gray-zone pairs require LLM inference—from under 5% of pairs with clean demographics to nearly all in heavily degraded data. In batch mode, MedGemma classifies 20 pairs/second on a datacenter GPU (H100), with sub-second per-pair latency even on consumer hardware (RTX 3090).

Integration. Input is standard HL7/FHIR demographic and clinical fields. Output is a deduplicated Master Patient Index—a drop-in replacement for existing MPI systems. Match and

Record A — Facility 1 (*Xuwo Haag*)

```
CONDITIONS:
  2023:  Medication review due
  2025:  Ischemic heart disease *
MEDICATIONS:
- Hydrochlorothiazide 25mg (2023-2026)
- Lisinopril 10mg (2023-2026)
OBSERVATIONS:
- Height:  159.6 cm | Weight:  70.7 kg
- A1c:  5.9% (2023-01), 5.9% (2025-10)
```

Record B — Facility 4 (*Xiao Ullrich*)

```
CONDITIONS:
  2025:  Medication review due; Stress *
  2026:  Limited social contact *
MEDICATIONS:
- Hydrochlorothiazide 25mg (2025-ongoing)
- Lisinopril 10mg (2025-ongoing)
OBSERVATIONS:
- Height:  159.6 cm | Weight:  70.7 kg
- A1c:  6.3% (2025-01), 6.2% (2026-01)
```

Figure 2: The hardest pair in the evaluation set: completely different names give Splink a match probability of just 0.0006. MedGemma identifies identical medications, matching biometrics, and overlapping care patterns to correctly classify this as a match with 99.99% confidence.

reject thresholds are configurable, allowing operators to tune the precision–recall trade-off for their clinical context.

User experience. For the ED registrar, ClinFuse means fewer fragmented charts—the MPI has already resolved ambiguous records using clinical evidence. For IT staff, the pipeline automates gray-zone resolution, producing a cleaner index requiring less manual reconciliation.

Confidence and auditability. Every resolved pair carries its decision source and confidence score, enabling operators to flag uncertain LLM decisions for human review and tune how aggressively ambiguous pairs route to clinical classification.

Open source and reproducible. The complete pipeline is implemented as a DVC directed acyclic graph with separate training and inference tracks. All stages run in Docker containers. Source code, model, adapter, and dataset are publicly available (see Acknowledgments).

Limitations and Future Work

Limitations. ClinFuse is evaluated on synthetic data (Synthea), which, while clinically realistic, may not capture all real-world patterns such as identity theft, deliberate data falsification, or extreme data sparsity in safety-net hospitals. The fusion weights are tuned for this dataset and would require calibration on real EHR data. Synthetic evaluation data may differ in difficulty from real-world data, so reported metrics demonstrate the approach's viability rather than guarantee production performance.

Future work. Key directions include: (1) validation on de-identified real EHR data to confirm generalization; (2) *tiered model escalation* where uncertain pairs route from smaller models to larger ones and ultimately to human-in-the-loop review, creating a graduated safety net for the most ambiguous cases; and (3) *multimodal resolution* leveraging MedGemma's vision capabilities to incorporate radiology images and scanned documents as additional identity signals.

Conclusion

ClinFuse shows that medical foundation models can serve as infrastructure components—not just diagnostic aids—when embedded in systems that invoke them selectively. If a 4B-parameter model fine-tuned for clinical QA can reduce record fragmentation by 73% on a task it was never designed for, the broader implication is that medical language understanding is a general-purpose capability applicable wherever clinical reasoning intersects operational systems: patient matching today, but also care-gap detection, formulary reconciliation, and the national-scale identity infrastructure that TEFCA and the MATCH IT Act demand.

Acknowledgments

This work uses MedGemma [8] (Google Health AI) for clinical classification and Splink [9] for probabilistic linkage. Synthetic data generated with Synthea [17]. Developed for the Kaggle MedGemma Impact Challenge [21].

Resources:

Code: https://github.com/abicyclerider/clinfuse
Model: https://huggingface.co/abicyclerider/medgemma-4b-entity-resolution-text-only
Dataset: https://huggingface.co/datasets/abicyclerider/entity-resolution-pairs
Video: *[URL TBD before submission]*

References

[1] ECRI Institute PSO, "Deep dive: Patient identification," 2016.
[2] G. Roda et al., "Double trouble: A propensity-matched cohort study evaluating the associations between duplicate medical records and patient outcomes," *BMJ Qual. Saf.*, 2026.
[3] B. H. Just et al., "Why patient matching is a challenge: Research on master patient index (MPI) data discrepancies in key identifying fields," *Perspect. HIM*, 2016.
[4] Black Book Market Research, "Provider interoperability and patient record error rates," 2018.
[5] ONC, "Patient identification and matching final report," 2014.
[6] U.S. Congress, "MATCH IT Act," H.R. 2002, 2025.
[7] ONC, "TEFCA v2.1," 2024.
[8] Google Health AI, "MedGemma model card," 2025.
[9] R. Linacre et al., "Splink: Probabilistic record linkage at scale," *Int. J. Pop. Data Sci.*, 2022.
[10] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *JASA*, 64(328):1183–1210, 1969.
[11] Z. Wu et al., "MedLink: De-identified patient record linkage," *Proc. KDD*, 2023.
[12] S. Gupta et al., "PRISM: Patient records interpretation for semantic matching using LLMs," *npj Digit. Med.*, 7, 2024.
[13] R. Peeters, A. Steiner, and C. Bizer, "Entity matching using large language models," *arXiv:2310.11244*, 2023.
[14] H. Li et al., "On leveraging large language models for enhancing entity resolution: A cost-efficient approach," *arXiv:2401.03426*, 2024.
[15] T. Wang et al., "Match, compare, or select? An investigation of large language models for entity matching," *Proc. COLING*, 2025.
[16] C. Cao et al., "Linking patient records at scale with a hybrid approach combining contrastive learning and deterministic rules," *medRxiv*, 2025.
[17] J. Walonoski et al., "Synthea: Generating synthetic patients and EHRs," *JAMIA*, 25(3):230–238, 2018.
[18] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.
[19] M. J. J. Bucher and M. Martini, "Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification," *arXiv:2406.08660*, 2024.
[20] F. Di Palo, P. Singhi, and B. Fadlallah, "Performance-guided LLM knowledge distillation for efficient text classification at scale," *arXiv:2411.05045*, 2024.
[21] F. Mahvar et al., "MedGemma Impact Challenge," Kaggle, 2026.